R E P O R T R E S U M E S

ED 020 138 SE 004 696

REPORT TO THE NATIONAL SCIENCE FOUNDATION ON THE SUBSTRUCTURE SEARCH DEMONSTRATION CONDUCTED IN NEW YORK CITY SEPTEMBER 1966.

AMERICAN CHEMICAL SOC., COLUMBUS, OHIO

PUB DATE 66

EDRS PRICE MF-$0.50 HC-$4.04 99P.

DESCRIPTORS- *CHEMISTRY, *COMPUTER ORIENTED PROGRAMS, *INFORMATION CENTERS, *INFORMATION STORAGE, *INFORMATION RETRIEVAL, *INFORMATION SCIENCE, *INFORMATION SERVICES, CHEMICAL ABSTRACTS SERVICE, AMERICAN CHEMICAL SOCIETY, NATIONAL SCIENCE FOUNDATION, SUBSTRUCTURE SEARCH SYSTEM,

CHEMICAL ABSTRACTS SERVICE (CAS), IN CONJUNCTION WITH THE NATIONAL SCIENCE FOUNDATION, CONDUCTED THE FIRST PUBLIC DEMONSTRATION OF CAS COMPUTER-BASED SUBSTRUCTURE SEARCH TECHNIQUES AT THE 152ND MEETING OF THE AMERICAN CHEMICAL SOCIETY IN NEW YORK CITY. FROM SEPTEMBER 11 THROUGH SEPTEMBER 16, 1966, INTERESTED PERSONS WERE GIVEN THE OPPORTUNITY TO SEE SUBSTRUCTURE SEARCH OPERATIONS AND TO DETERMINE THE TECHNIQUES, CAPABILITIES, AND POTENTIALITIES IN LIGHT OF THEIR OWN NEEDS. THE PURPOSE OF SUBSTRUCTURE SEARCHING IS TO ENABLE TECHNICAL PERSONNEL TO AUTOMATICALLY SEARCH FOR CHEMICAL STRUCTURES AND SUBSTRUCTURES THAT HAVE BEEN REPORTED IN THE LITERATURE AND REGISTERED IN THE CAS CHEMICAL COMPOUND REGISTRY SYSTEM. THE NEW YORK CITY DEMONSTRATION USED A "BREADBOARD MODEL" OF THE SUBSTRUCTURE SEARCH SYSTEM. THIS IS A VERSION CAPABLE OF PRODUCING ALL OF THE CORRECT ANSWERS TO THE QUESTIONS, BUT A MODEL WITHOUT THE REFINEMENTS THAT WILL BE AVAILABLE IN AN OPERATIONAL SYSTEM. BECAUSE OF THE DIALOGUE BETWEEN CAS SCIENTISTS AND PRACTICING CHEMISTS AND CHEMICAL ENGINEERS, CAS IS NOW ABLE TO MAKE SIGNIFICANT TECHNICAL IMPROVEMENTS TO BETTER SERVE THE INFORMATIONAL NEEDS OF THE CHEMICAL COMMUNITY. THIS REPORT DESCRIBES THE DEMONSTRATION, THE BREADBOARD MODEL, AND THE RESULTS OF THE DEMONSTRATION, AS WELL AS THE IMPROVEMENTS SUGGESTED AS A RESULT OF THE EXPERIMENT. (DS)

ED020138

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

REPORT TO THE NATIONAL SCIENCE FOUNDATION

on the

SUBSTRUCTURE SEARCH DEMONSTRATION CONDUCTED IN NEW YORK CITY SEPTEMBER 1966

CHEMICAL ABSTRACTS SERVICE

AMERICAN CHEMICAL SOCIETY

SE004 696

Copy No. 48

Report

to the

NATIONAL SCIENCE FOUNDATION

on the

SUBSTRUCTURE SEARCH DEMONSTRATION

Conducted in New York City

September 1966

CHEMICAL ABSTRACTS SERVICE

AMERICAN CHEMICAL SOCIETY

CONTENTS

	Page
ABSTRACT	1
DEMONSTRATION OBJECTIVES	2
THE CAS SUBSTRUCTURE SEARCH SYSTEM	4
DEMONSTRATION DESCRIPTION	9
DEMONSTRATION RESULTS	14
SYSTEM ENHANCEMENT BASED UPON DEMONSTRATION RESULTS	22
STATISTICAL SUMMARY AND COSTS	27

APPENDIXES

A. Glossary
B. Screening and Iterative Search Data
C. Screens
D. Demonstration File Characteristics
E. List of Questioners and Their Affiliations
F. Examples of Questions and Retrieved Answers

FIGURES

Page

FIGURE I
Typical Substructure Search Question 6

FIGURE II
Search System 10

FIGURE III
Information Flow 11

TABLES

Page

TABLE I
Question/Hit Statistics 28

TABLE II
Computer Times 29

TABLE III
Substructure Search Computer Cost Analysis 32

ABSTRACT

Chemical Abstracts Service (CAS), in conjunction with the National Science Foundation, conducted the first public demonstration of CAS computer-based substructure search techniques at the 152nd Meeting of the American Chemical Society in New York City. From September 11 through September 16, 1966, interested persons were given the opportunity to see substructure search operations and to determine the techniques capabilities and potentialities in light of their own needs.

The purpose of substructure searching is to enable technical personnel to automatically search for chemical structures and substructures that have been reported in the literature and registered in the CAS Chemical Compound Registry System. The New York City demonstration used a "breadboard model" of the Substructure Search System, i.e. a version capable of producing all of the correct answers to the questions, but a model without the refinements that will be available in an operational system. Because of the dialogue between CAS scientists and practicing chemists and chemical engineers, CAS is now able to make significant technical improvements to better serve the informational needs of the chemical community.

This report describes the demonstration, the breadboard model, and the results of the demonstration, as well as the improvements suggested as a result of the experiment.

DEMONSTRATION OBJECTIVES

The New York demonstration was an experiment. It was designed to determine, under conditions approaching those that could be expected to prevail for an operational system, the adequacy and efficiency of the substructure search techniques and their reception among practicing chemists, chemical engineers, and others who require chemical information. Among the many specific objectives of the demonstration were the following:

1. To acquaint the technical public with a computer-based technique that would rapidly recall and collate chemical data based on chemical structures, and to allow the CAS staff to gain valuable experience in such areas as question framing and coding, dialogue with users, and remote-location operations.
2. To determine the types of questions that would be asked and, in general, to determine what the practicing chemist and chemical engineer wanted the system to do for him.
3. To assess existing techniques for such procedures as screening, coding, and remote searching, and to collect additional design data that might lead to their improvement.
4. To acquire actual operating data such as machine times and answers per question.

The success of the demonstration in meeting the goals outlined above is summarized in the following sections. A glossary of terms used in

substructure searching appears in Appendix A, while detailed statistical data on the questions asked and answers retrieved are provided in Appendix B. Appendix C gives detailed information about the screens used. Appendix D gives characteristics of the Demonstration File, Appendix E lists the questioners and their organizational affiliations, and Appendix F provides examples of the questions asked and hits retrieved.

THE CAS SUBSTRUCTURE SEARCH SYSTEM

The Substructure Search System is being developed as part of the overall CAS computer-based chemical compound-handling system. It is being designed to locate within a file of structures in connection table form all compounds that possess one or more specified substructures.

Essential to the concept of the Substructure Search System is the provision of maximum flexibility in both question and answer specificity. To provide the desired flexibility, the search technique being designed at CAS operates at several levels of specificity. At one level, chemical fragment screens, many of which correspond to functional groups with which every chemist is familiar, are used to select from the whole file those compounds that include potential answers to the question. Such screening is a very rapid and relatively inexpensive way to select compounds from a file. Depending upon such things as the size of the list of answers, the relationship between the sought-after substructure and the retrieved structure, and the cost of the search, this level of search may provide answers that are quite satisfactory to the questioner. Nevertheless, if greater specificity is desired, an iterative, atom-by-atom, bond-by-bond search level is available which can reduce the list of candidate structures to include only those that meet the more exact specifications. In no case will the search system eliminate exact answers to a question--rather, the "non-answers" are rejected. At each level of specificity, the user will have the option to either terminate or continue the search, based upon the results of the previous step.

Clearly, screening is a critical phase of substructure searching, one that determines the efficiency of the search and hence the cost of searching. Most screens are produced by relatively inexpensive screen generation programs that automatically strip the various fragment types (atom counts, ring counts, etc.) from the computer structural record. Approximately 1500 such fragments are used for screening purposes.

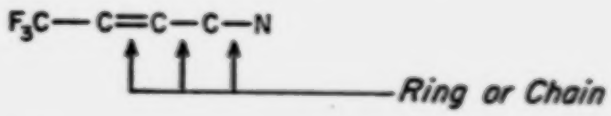
Screens are represented in the computer file by "bit indicators." Associated with each compound in the file is a series of these indicators (binary digits), each of which corresponds to a specific screen item. Each bit acts like a switch: if the compound possesses the screen item corresponding to a particular bit, the bit is set to "on." If the compound does not possess that item, the bit remains off. Once such a record has been established for each compound in the file, screening can be accomplished for a substructure search question merely by setting up a bit-indicator record for the question showing the screen items to be located. The record for the question is then compared with the records for the compounds on file in a quick and easily accomplished computer procedure. It should be noted that once the indicators are set for a file of compounds it is not necessary that they remain static. These screen assignments can be altered to fit a given operating environment.

The Substructure Search System incorporates Boolean logic, and questions may be posed in terms of "and", "or", and "not" logic. "And" logic requires the presence of an atom or group of atoms in the answer. "Not" logic specifies that an atom or group of atoms must not be present in the answer. "Or" allows alternatives, one of which must occur in every retrieved structure.

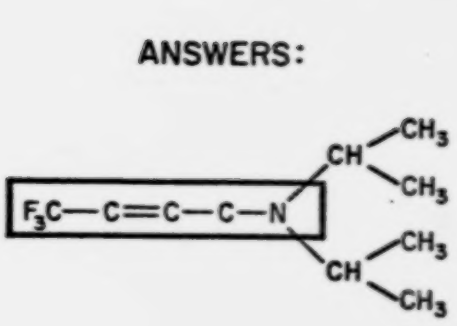
A fourth listing, "Don't Care", allows atoms and bonds within the substructure to be left unspecified.

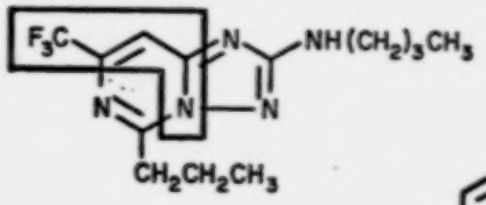
Figure 1 shows a typical substructure search question and illustrates how answers are dependent upon question specificity. The question allows the three bonds marked by arrows to appear in either a ring or a chain.

TYPICAL SUBSTRUCTURE SEARCH QUESTION



ANSWERS:





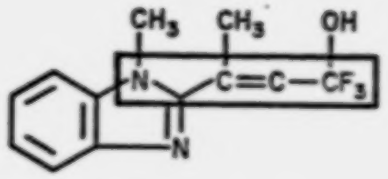


FIGURE 1

The first answer shows the substructure imbedded within two rings. The second answer has no rings, while the third is a ring-chain combination. Had the "don't care" bonds of the question all been limited to ring bonds in the indicated positions, only the first answer would have been satisfactory. Had they been limited to chain bonds, only the second answer would have been obtained. In neither of these last two possibilities would the third structure have been retrieved.

The CAS Substructure Search System is experimental, and certainly not all of its potential uses have even been recognized. Therefore, it is expected that many more than the four applications outlined below will be

found for the system as it matures and as potential users become more acquainted with it.

1. The general use to which the system will be put is that of substructure search. That is, searches of a file of structures to locate those that contain similar structural characteristics. Such searches are by no means limited to the CAS computer, they could be conducted by other institutions or organizations.
2. Since compounds containing specified substructures can be identified during the registration process, this system can provide an alerting service for new compounds containing substructures of interest to any given user. Moreover, since all ring systems indexed in the subject index to Chemical Abstracts are registered, any new ring system entering the system can automatically be identified, even when it is embedded in another structure.
3. The system provides the mechanism for automatically generating fragmentation codes for updating a user's fragment search file whether it be computerized or manual. By interrelating the fragments of a manual system and the screens of an associated computer search system, the latter can be used efficiently to supply more specific answers than are obtainable by a manual search.
4. If a substructural hierarchy is established--which may be varied at each use--for printing out a list of answers, the system can be used to organize a series of structures without depending upon systematic nomenclature or human intervention. In addition, if structures are available directly from the computer, it is

possible to pose whole-question or substructure questions directly to the system in diagrammatic language and receive an organized list of answers in the same form. This work has already been accomplished for small systems by several groups and is now under development for large systems by CAS.

DEMONSTRATION DESCRIPTION

The substructure search techniques demonstrated at New York City were performed using a "breadboard" model of the operational system. That is, the components used for the demonstration were not specifically designed to be interlinked, and although the demonstrated system was fully capable of selecting all of the answers from the files for a substructure search question, it did not possess the operational sophistication required of a heavily used system--it did not perform many of its functions in an efficient manner. Moreover, some of the tasks that will eventually be performed, partly or entirely, by computer were assigned to humans for the demonstration, and some of the options that will eventually be offered routinely were available only by dividing questions into two or more parts. Finally, the demonstrated system was programmed for the IBM 7010 computer, whereas the first operational system will employ the IBM 360 computer. Nevertheless, the system was fully capable of its basic task--computer searching for defined substructures in a file of compound-structure representations.

To limit the amount of time and money spent to search each question, while at the same time providing representative search results, CAS set up a special demonstration file of 55,396 compounds; slightly more than one-tenth of the number of compounds registered as of September 1966. (See Appendix D for a description of this file.) In addition, the number of hits provided to any one question was arbitrarily limited to six.

A generalized flow diagram of the Substructure Search System is shown in Fig. 2, while Fig. 3 details the "System" as it operated specifically

SEARCH SYSTEM

Substructure Search Demonstration
New York City, September, 1966

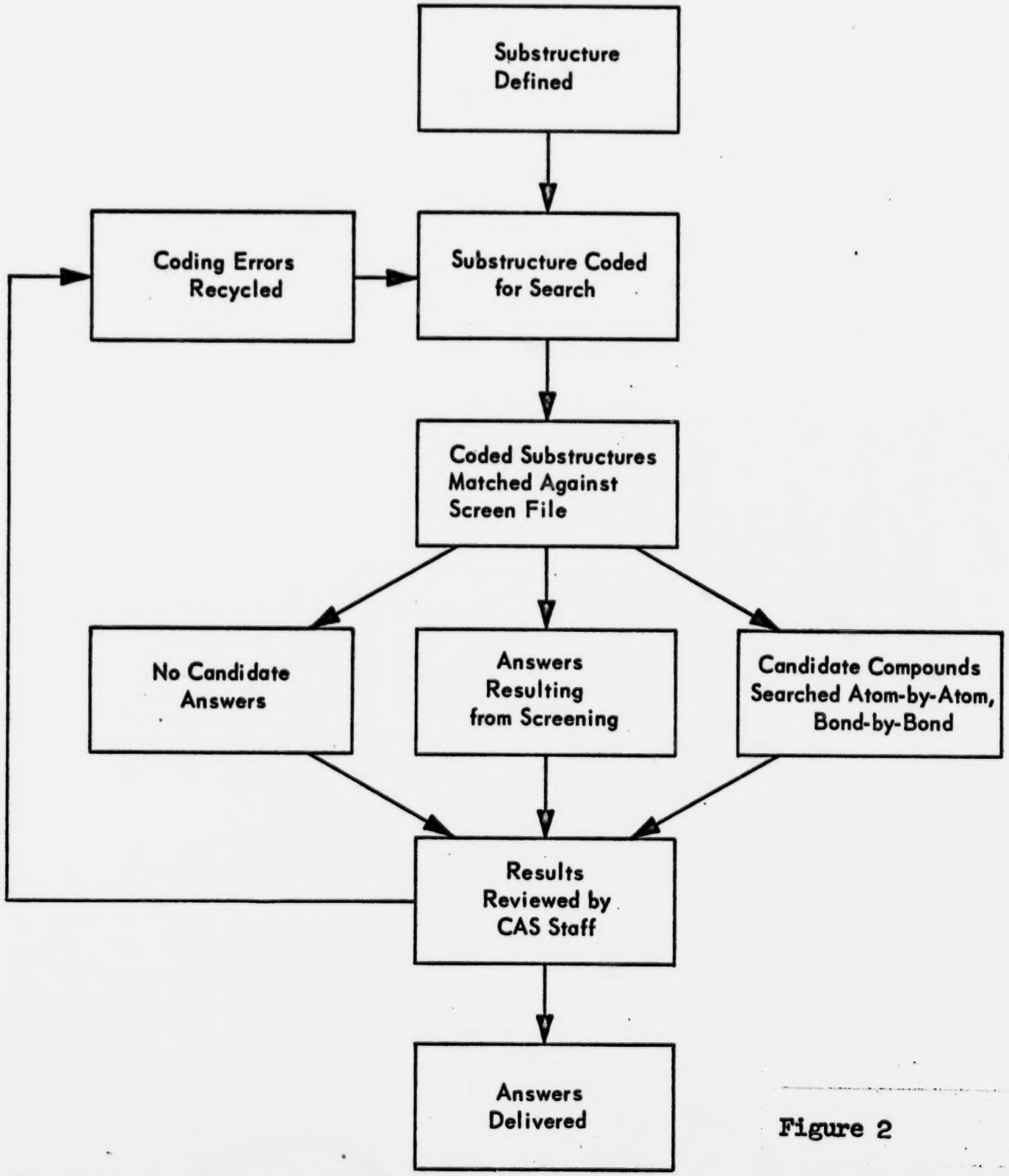


Figure 2

INFORMATION FLOW

Substructure Search Demonstration
New York City, September, 1966

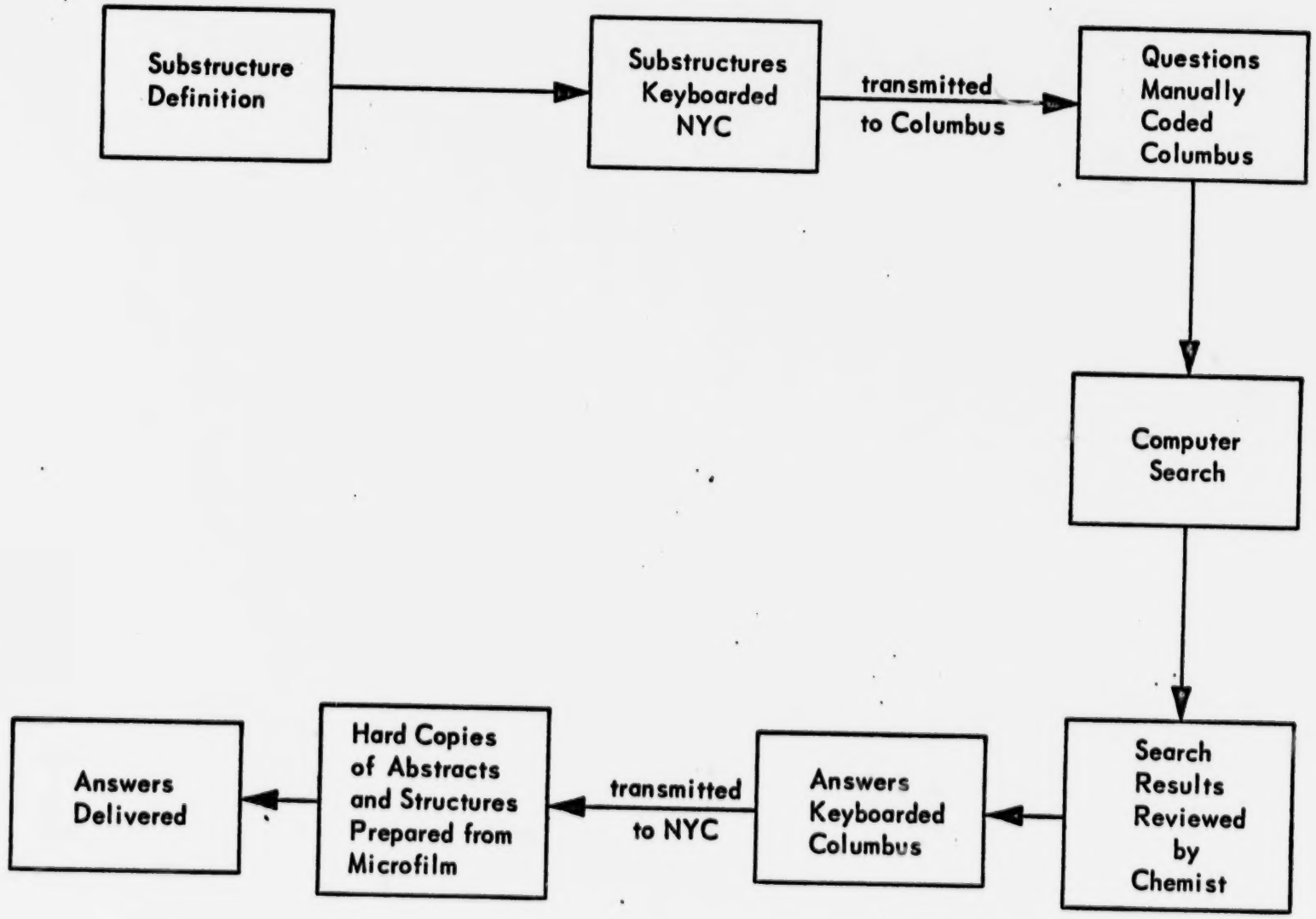


Figure 3

for the New York City demonstration. The process started with a face-to-face interview between a questioner and a CAS chemist to determine the exact details and precise meaning of a question and the objectives of the search. Once these were discerned, the substructure was drawn and the question keyboarded on a paper-tape-generating structure typewriter located in New York. The data contained on the paper tape generated by the typewriter were then transmitted by TWX to CAS headquarters in Columbus, Ohio where a hard copy was produced by a similar typewriter.

In Columbus, the screens were coded manually by a CAS chemist. The coded substructure search questions were then matched against the Search Screen File--a file which included only the Bit Indicator Screens for each compound on the Search File and the corresponding Registry Numbers. This screening process produced a set of Registry Numbers as candidate compounds. At this point, some of the questions were completely answered because the screening process determined that no exact answers existed on file or because screening completely identified the exact answers. For the other questions, the corresponding sets of candidates included not only those compounds that exactly answer the search question, but also some related compounds. For the latter sets, an iterative search, atom-by-atom and bond-by-bond, was made on the candidate compounds in the Structure File to select the exact answers, referred to as "hits", to the search question.

At this point, the structure of the compound, the molecular formula, and the CA index name were reviewed by a chemist to insure that the results were valid. Errors in coding were then cycled for recoding and re-search.

The validated structure and bibliographic data were then typed on the structure typewriter and transmitted to New York where hard copies of the information were produced on the structure typewriter located there. The answers were then sorted and the abstract for one bibliographic citation retrieved from CA on microfilm included with the printed answers. These and a copy of the search question were later returned to the questioner.

DEMONSTRATION RESULTS

Substructure searching has been in the development stage for several years, but until the time of the New York City demonstration, the capability had never been shown publicly.* CAS believed that if the operational system was to accomplish its goal--to fill a major need in the chemical researcher's information requirements--the existing system required public exposure. The New York meeting gave us such an opportunity. Through a special exhibit set up at the ACS meeting, some 750 people were introduced to the search technique. These people were provided with literature on substructure searching and had the opportunity to discuss the system with CAS staff and to test the system by supplying questions to it.

Some 163 persons representing approximately 110 organizations--universities, industrial firms, governmental agencies, and research institutes--availed themselves of the opportunity to ask questions, and 183 searches were run during the four-day demonstration. About half of the questioners were research chemists, while the other half were chemical information specialists. Appendix E lists the questioners and their affiliations.

To provide experience to its staff, CAS assigned eleven chemists and six systems personnel to conduct the New York demonstration. Five chemists and three systems personnel were located in New York, the remainder in Columbus. This staff was aided by a chemical-typewriter operator at each location as well as keypunch operators in Columbus.

---

*CAS did demonstrate an earlier version for government representatives in November 1965.

Since this demonstration was to be our first experience in handling a large and widely diversified number of substructure search questions, each professional involved underwent approximately 20 hours of training prior to the meeting. Items such as the following were discussed to familiarize those involved with the skills they would need:

a. Interaction with questioners.

b. Problems of question definition.

c. Problems of communications between New York and Columbus

d. Coding for screens and iterative search.

The Pattern of Questions and the Requested System Capabilities

The New York demonstration gave CAS an opportunity to gather information as to the types of queries that could be expected to be asked of an operational system and to identify specific system characteristics desired by users.

Discussions between CAS personnel and visitors to the demonstration made it clear that any operational system must be flexible enough to serve the spectrum of users, from the single researcher working at a university to the research section of a large industrial firm. Question profiles, search files, answer specificity, and output formats each pose special problems that will vary according to the environment in which the system is used. It is also clear from the demonstration that the system must be capable of performing searches for both full structures and substructures, retrospectively and on a current-awareness basis.

Even though it will be necessary that the system be customized in terms of such items as the types of acceptable questions and the format and detail

of output, individual queries will have to be exactingly defined. Each bond, each element, each alternative must be precisely identified--even if it is "don't care"--if the user is to obtain the response he requires. In New York, CAS personnel questioned the user extensively to obtain this information. Generally, we found that, although his questions were very specific, they were imprecisely worded, and it required considerable time to define the inquiry with sufficient detail to insure that the questioner would receive the answers he desired. Such personal interrogation will not ordinarily be available in a highly automated system. Instead of face-to-face interrogation, it is expected that, in an operational system, the computer will ask the pertinent questions that will lead to fully defined structural questions. Computer-user dialog will help both the novice and the experienced user to obtain satisfactory results from the system with a minimum of effort.

Registry Numbers alone will be of little value in most applications. As a minimum, users will have to be provided with Desktop Analysis Tools or direct computer output that links the Registry Numbers with names that can be searched for in printed indexes and/or structural formulas. A range of output options must be provided; the user will want bibliographic citations, titles, structures, and/or other printed information to help him determine the references that contain relevant information. Perhaps hard copies of abstracts or even the actual articles may be included as part of the package. The choices that will be ultimately available to the consumer are heavily dependent upon other systems and services being developed at CAS and elsewhere.

In talking to individuals, CAS staff received several requests that pointed to capabilities in the overall structure-handling system that needed strengthening. These include below:

(1) the capability to integrate full-structure, substructure, and nomenclature searches without the need for the user to make a distinction between the various systems.

(2) the capability of searching for compounds containing specified isotopes.

(3) the capability of allowing the user to stipulate that certain compounds containing the sought-after substructure will be excluded from the answers.

(4) the capability of searching for substructures that contain a repeating group (polymeric or not) attached to a specified group at each end, without specifying the number of repetitions.

(5) the capability of searching for structural information on polymers and coordination compounds.

(6) the general capability of making correlative searches (with appropriate logic) that utilize both text materials and structural information, whether from the same file or from interrelated files.

All of the above suggestions are based on specific questions for which the above capabilities could have been utilized. For example, some questioners were interested in compounds possessing a specific substructure, but they did not want to recall the compounds that they already knew contained that substructure. Although these results can be achieved by manual screening of the answers, CAS believes that in the interest of economy and

accuracy, the ability to exclude predetermined information should be included in the system.

Screening, Coding, and Remote-Searching Techniques

Because of its importance to the total system, the screening program of the Substructure Search System has received continuing attention and its efficiency has steadily been improved.

An important objective of the New York demonstration was to determine screening efficacy as a function of the questions asked. To make such an evaluation, the concept of "percent screenout" is used and defined as:*

$$\frac{\text{No. of Compds. Eliminated by Screening}}{\text{Total Compds. in File}} \times 100$$

Applying this criterion, we find that of the 183 questions asked, 67% (123) questions were screened with at least 99% efficiency. That is, 1% or less of the compounds on file for the demonstration passed the screens. Thirty-six, or 20% of the questions were screened with 95-99% effectiveness, 6% with 90-94% effectiveness, and 6.5% were screened with less than 90% effectiveness.

*This criterion is useful on the assumption that the number of answers to a substructure search question will be a very small percentage of the total file. Under these circumstances, a screenout percentage near 100% indicates effective screening. However, for a question in which the number of answers is a significant percentage of the file size, the percent screenout will be small even when screens operate perfectly such that no iterative searching is required. For example, one question asked for during the demonstration had 14,806 answers, all of which were found by screening. Yet the screenout percentage was only 73.3% for this question, since the total file size was 55,396.

Of the questions with less than 99% screenout, several had more than 550* answers (1% of the file), and therefore could not have realized 99% screenout. Nevertheless, other questions for which large numbers of compounds passed the screens highlighted the need for some additional screens. Several that are to be added to the system are:

(1) screens for carbocyclic and heterocyclic rings of specified sizes;

(2) generic level screens for a carbocyclic ring of any size and for a heterocyclic ring of any size;

(3) a carefully selected group of screens for complex ring systems such as those illustrated by anthracene, phenanthrene, and benzindene, etc.;

(4) additional chemically significant fragments, including some for atom chains of varying lengths (e.g., 4, 5, or 6 atoms);

(5) addition of some generic-level, chemically significant fragments which would simply show connectivity relationships without specifying particular atoms.

At present, most of the screens used in substructure searching are structurally specific--they require the presence of specific atoms, specific bonds, etc.--for all potential answers to the substructure search questions (see Appendix C for screen descriptions). Consequently, the less specific the search questions (e.g., the greater the number of "don't care" atoms or bonds), the less effective are the screens. A few generic level screens were used for the New York demonstration, but our experience there taught

*Projected figures. Only six answers per question were provided during the demonstration.

us that a substantially greater number are required for effective screening. The screens developed to fulfill the needs described in Nos. 4 and 5 above are described in the next section of this report.

Another technique that was evaluated in light of the New York demonstration was the coding of substructure search questions. Our experience at the demonstration bore out our previous feeling that the human encoding of questions now required is too inefficient to be used in an operating system. The human requirements for coding are too extensive to be performed for any system subject to heavy use. In addition, manual coding is too complex to handle without extensive training. Although the need for some limited amount of manual coding of questions, for both screening and iterative search, may always exist, it must be simplified. However, it is expected that in the operational system, the computer will be the major instrument used to code questions for both screening and iterative searching. The coding procedure will probably be started as the user types the structure on a chemical typewriter or possibly on an on-line device such as the IBM 2250 (essentially a chemical typewriter incorporating a cathode-ray tube with a light pen for real-time playback). Through a translation program such as now used for the Registry System, the information will be coded to a connection table from which the screens will be generated. Through another translation program, the information will be coded in the form needed for the iterative search. Throughout this process, a computer-directed dialogue with the user will help him to frame his question with appropriate precision to maximize his ability to gain useful answers from the system.

A related question investigated in light of the New York City demonstration involved the techniques of remote searching and the handling of structures on the structure-typing typewriter.* In general, both procedures were entirely satisfactory, the structure typewriter proving a useful tool, and the remote terminal setup operating satisfactorily. Although about a dozen questions were either garbled in transmission or were typed at the remote end with insufficient information; this caused only minor problems that were solved with a telephone call.

*Modified Dura Mach 10. For a detailed description of this instrument, refer to "Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures", J. M. Muller, Journal of Chemical Documentation, Vol. 7, No. 2, pp 88-93.

SYSTEM ENHANCEMENT BASED UPON DEMONSTRATION RESULTS

One of the major reasons for the New York demonstration was to give CAS an opportunity to detect areas within the system that needed further investigation. Because the substructure searching methodology was subjected to a critical review by the type of individuals that would use an operational system, we were able to detect the areas that needed strengthening. Two such areas are discussed below.

Additional Screen Capability

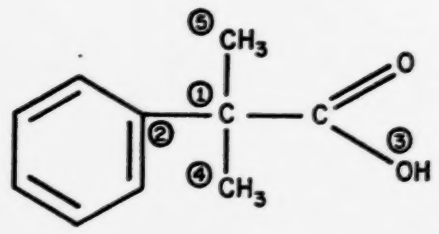
It has been previously stated that one of the more important lessons learned from the New York demonstration was that it pointed to the need for additional screen types, including some intermediate generic screening capabilities. Concerning the latter, most of the approximately 1500 screens used for the demonstration were either too specific or so generic so as to reduce the screen efficiency below acceptable limits for certain questions. Had certain screens of intermediate generic nature been available to rapidly separate the candidate structures from the total file, less iterative searching would have been required. In an operating system, this would result in a less expensive operation since, as would be expected, screening is much less time consuming than iterative searching.

As a result of the demonstration, two new types of screens are being instituted in the Substructure Search System. These are (1) the Degree of Connectivity Screen and (2) the Linear Sequence Screen. In addition, some

intermediate generic levels are being introduced into the "Triplet" and "Moiety" screens.

1. Degree of Connectivity Screen

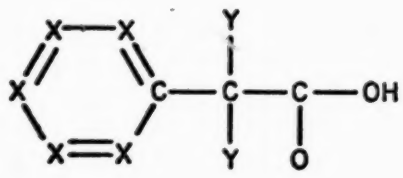
This screen type is defined as the minimum number of atoms having N or more nonhydrogen attachments, where N can equal 3, 4, 5, 6. For example, the structure illustrated below would satisfy the substructure search requirement of possessing one or more atoms with a degree of connectivity of four, because Atom No. 1 has four nonhydrogen atoms attached to it (Nos. 2, 3, 4, and 5).



The structure would also satisfy the requirements for three or more atoms with a degree of connectivity of three, since Atom Nos. 1, 2, and 3 each have at least three nonhydrogen atoms attached.

The advantage of this screen type is that it enables one to utilize the discriminatory power of atoms having degrees of connectivity of 3 or greater even when all of the atoms in the substructure search request are not identified specifically. For example, the substructure search request illustrated below has one atom with a degree of connectivity of 4 and also has three atoms with a degree of connectivity of 3 or more (the atom with a degree of connectivity of 4 is also counted as an atom with a degree of

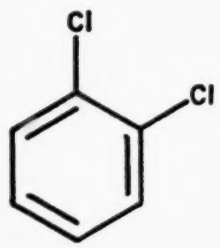
connectivity of 3 or more). The use of these screens in conjunction with others that are now available will help reduce the number of structures that will have to be searched atom-by-atom.



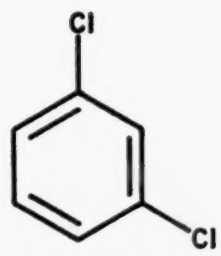
X = C or N
Y ≠ H

2. Linear Sequence Screen

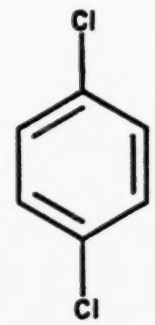
The Linear Sequence Screen is defined as a series of 4, 5, or 6 specific atoms and the bonds uniting them. The only bond specificity is whether they are chain or ring bonds. For example, among the linear sequences present in the three structures shown below, the most discriminatory ones are indicated beneath each structure (a ring bond is designated by an asterisk and a chain bond by a hyphen).



Cl–C*C–Cl



Cl–C*C*C–Cl



Cl–C*C*C*C–Cl

Such discriminating power was not possible with our previous screens. Consequently, in a structure search request for ortho dichlorobenzene all three of the above structures would have passed the screens and would have to be iteratively searched. With the Linear Sequence Screen described above, two of the structures would be screened out, thereby reducing the total amount of iterative search time required to retrieve the desired structures.

Computer Editing of Search Questions

The experimental substructure search "system" demonstrated in New York incorporated a number of computer editing routines to check the validity of information coded in the screens for each question. Among these were checks for keyboarding errors and checks to substantiate that coded screens were available for use. However, since completely computerized editing routines had not as yet been provided, a substantial amount of manual editing had to be done--far more than would be tolerable in an operational system. To reduce the amount of manual effort required for this purpose, appropriate computer editing routines will be written to check the validity of information coded for search. Examples of the type of editing checks to be provided are:

1. Checks for Allowable Characters in a Given Column

In the search coding operations, the type of character allowed in certain columns is restricted. For example, only numeric characters are allowed in the columns of bond values and valences. In the operational system, if an alphabetic character is mispunched in one of these columns, the information will be rejected and appropriate diagnostics describing the reason will be produced. In other instances, only certain characters may be entered. For example, in the columns reserved for Boolean Logic operators, only the letters A, O, or N (for "AND", "OR", and "NOT" logic, respectively) are allowed. If an invalid character is used, the information will be rejected and an appropriate diagnostic will be produced.

2. Checks for Allowable Data in Certain Fields

The type of data allowed in certain fields is also restricted. For example, several two-column fields will accept only the symbols for the elements or certain numerical values assigned to atoms that have been previously cited elsewhere in the iterative search question. If invalid data (e.g., an invalid element symbol, or an invalid number) appears in one of these fields, the information will be rejected and an appropriate diagnostic will be produced.

STATISTICAL SUMMARY AND COSTS

This section of the report presents a summary of statistics, including costs, that were derived from the New York demonstration. More detailed statistics concerning screening and iterative search can be found in Appendix B.

Throughout this section and the succeeding appendixes, the term "hit" is used. This term is defined to mean a structure retrieved by a search that exactly satisfies the search question. The term is used to contrast answers that identify structures (hits), and the situation in which a search produces no Registry Numbers of structures that satisfy the question because none exist in the file. Although this latter circumstance is not defined as a hit, it nevertheless is a valuable piece of information.

It should also be recalled that because of time and cost limitations, an arbitrary limit of six hits per search was imposed. That is, once six hits were retrieved for a question during a search, the search was terminated.

TABLE I

QUESTION/HIT STATISTICS

1.	Number of compounds on file to be searched	55,396
2.	Number of persons asking questions	163
3.	Number of questions asked	183
4.	Number of negative (i.e., no-hit) searches	81
5.	Number of searches with hits	102
6.	Number of hits (max. of 6/search)	529
7.	Projected number of hits[1]	35,745
8.	Range of number of projected[2] hits per search (based on 102 searches)	
	a. Maximum	14,806
	b. Minimum	1

---

[1] Projected figures estimate number of hits if the limit of 6 hits per question were not imposed.

[2] Because of the 6-hit limit, not all compounds passing screens were searched atom-by-atom. Instead, atom-by-atom search continued only until 6 hits resulted.

TABLE II

COMPUTER TIMES

	Actual (limit-6 hits/question)	Projected[1] (no limit)
1. Total Searching Times		
a. Screening	6.96 hrs.	6.96 hrs.
b. Iterative search	12.44	47.40
c. Total search time	19.40	54.36
2. Average Searching Times		
a. Screening time per question (based on 183 questions)	0.038	0.038
b. Iterative search time per question	0.068	0.259
c. Search time per question	0.106	0.297
d. Search time per hit	0.037[2]	0.0016[3]

(1) Only iterative search times are affected by the limit of six hits per question since screening is performed on the entire file while iterative search was terminated after six hits were retrieved.

(2) Based on the 529 hits actually retrieved.

(3) Based on a projected number of hits of 35,745.

Demonstration Costs

The following table compares the search costs incurred during the New York demonstration to those incurred during the demonstration held at CAS in November, 1965. This table should be used carefully since there were differences between the two demonstrations that have an affect upon the results. These differences are:

1. Although the search file was identical for both demonstrations, the screens used were not. The screens used for the earlier demonstrations were not all a part of the bit indicator record. Therefore, screening was much less efficient.
2. There was a limit of 15 hits per search question placed on iterative (atom-by-atom, bond-by-bond) search during the earlier demonstration, however, only one search was affected. A limit of six hits was placed upon the number of hits per search during the New York demonstration, and several questions were affected by the limit.
3. During the earlier demonstration, 25 questions were asked. Of these, 20 were iteratively searched.
4. The earlier demonstration used an IBM 1410 computer while the 1966 demonstration used an IBM 7010 computer.

Besides being able to compare this demonstration with the one that took place earlier, we are able to make some limited judgements as to costs that will be incurred when the 360 Substructure Search System becomes operational. Based upon preliminary cost estimates, a reduction of some 60% in search costs is anticipated at that time. This is, in part, due to the increased

speed of the computer and, in part, due to the shift from a breadboard to a designed search system. Based upon the 183 questions and the 529 hits developed at the demonstration, the cost per question for the operational system will be approximately $15.50 and the cost per answer, $.08 as compared to $38.61 per question and $0.21 per answer for the demonstration, based upon the projected figures. Neither set of figures includes the cost of generating the screen file since the allocation of such costs is dependent upon the size and character of the search file and the number of questions to be run against the file during its length of useful life.

TABLE III

SUBSTRUCTURE SEARCH

Computer Cost Analysis[1]

Demonstration Date & Place	Computer Used[2]	Limitations	Screening Per Search		Atom-By-Atom Search Per Search Question		Total Search Per Search		Total Search Per Hit	
			Hrs.	Dollars	Hrs.	Dollars	Hrs.	Dollars	Hrs.	Dollars
Columbus, Ohio November, 1965	1410	15 hits/ search 25 questions 2076 hits	.11	14.30	.34 (3)	44.20	.45	58.50	.0048	0.62
New York City September, 1966	7010	6 hits/ search 183 questions	.038	4.94	.068 (4)	8.84	.106	13.78	.037	4.81
		projected 183 searches 35,745 hits	.038	4.94	.259	33.67	.297	38.61	.0016	0.21

Notes:

(1) 55,396 compounds on file-cost of screen generation not included
(2) All computer time costed at $130/hr.
(3) 20 questions iteratively searched
(4) 75,443 iterative searches

APPENDIXES

APPENDIX A

GLOSSARY

GLOSSARY

Bit Indicator Screen--A screen in which a series of binary digits (bits) are each assigned a yes-no relationship for the presence of a given structural feature in a given compound.

Chemical Fragment--A well-defined grouping of atoms and bonds thought of as an entity, and from which bonding to other elements may or may not be well defined.

Connection Table--A computer-based linear notation consisting of atom-by-atom, bond-by-bond inventory that shows each atom, the atoms connected directly to it, and the types of linking bonds. Mass number, coordination number, valence, and charges are shown whenever they are required for exact identification. (Stereochemical data are included but were not machine searchable during this demonstration.) The connection table is comprised of the F-1, F-2, F-3, and F-4 records.

F-1 Record - That portion of the connection table that describes only the graph of the corresponding structural diagram; that is, only the connection network without specifying the character of the bonds (i.e., line values) or the node identities (i.e., the element symbols corresponding to the atoms in the diagram). Hydrogen atoms are not included in the definition of the graph.

F-2 Record - That portion of the connection table that identifies the type of atom corresponding to each network node appearing in the graph of the F-1 record.

F-3 Record - That portion of the connection table that specifies the bonding character of each line appearing in the graph of the F-1 record.

F-4 Record - That portion of the connection table that describes the qualifiers of the two-dimensional structure diagram. This portion of the record contains stereochemical descriptors, non-routine valence, isotopic number, hydrogen-atom count, and Registry Number.

Iterative Search--An atom-by-atom, bond-by-bond comparison between the substructure defined in a question and the connection tables on file. This process provides only exact answers to search questions.

Moiety--A chemical fragment.

Percent Screenout--Percent screenout is defined by the mathematical expression:

$$\frac{\text{No. of Compds Eliminated by Screening}}{\text{Total No. of Compds. in File}} \times 100$$

Question Coding--The translation process required to convert a search question into the symbolic form required by the computer program.

Registration--The process of determining the existence or absence of a substance in the Registry Files. The process includes the assignment of a Registry Number (see below) to each substance that is new to the files.

Registry Number--The unique nine digit (the ninth digit is a computer-calculated check digit) number which is assigned to each substance when it first enters the Registry and which is recalled each time that substance is checked against the file. The Registry Number may be used to identify

fully the substance, and it is used as the address in specialized subject files to identify data associated with the substance. In the Registry System, a Registry Number also is the file address for bibliographic and nomenclature data related to the corresponding compounds.

Registry System--The interrelated set of files directly associated with registration and the processes for accomplishing registration. These computer files include structural records, the molecular formulas, nomenclature, and bibliographic data.

Screen--A common structural characteristic identified in the search files as part of the corresponding structural diagram. The individual screens are selected partly on the basis of the frequency with which they appear as a part of a substructure search question, and partly on the basis of the frequency with which they appear in the search file. In the search system, a set of screens amounts to a conveniently arranged series of yes-no answers to commonly asked substructure search questions.

Screen Dictionary--A listing that defines the screens available for substructure searching.

Substructure--A specified set of atoms interconnected in a specified way; this constellation normally represents less than a complete molecule. (cf. Chemical Fragment)

APPENDIX B

SCREENING AND

ITERATIVE SEARCH DATA

SCREENING AND ITERATIVE SEARCH DATA

This appendix contains both detailed and summary data concerning the screening and iterative searches conducted during the demonstration. Table I, the Summary, presents quartile figures where appropriate to enable the reader to better evaluate the spread of various data. Table II presents the raw data as a function of each individual search question.

TABLE B-I

SUMMARY OF SCREENING AND ITERATIVE SEARCH DATA

	Number	First Quartile	Second Quartile	Third Quartile	Fourth Quartile
1. Number of compounds screened out per question	5,766 - 55,396				
2. Total number of compounds requiring iterative search (if no limit had been imposed)	312,557				
3. Number of compounds iteratively searched (limit of six per search)	36,942	2 - 12	13 - 63	69 - 244	283 - 6297
4. Number of compounds per question requiring iterative search (if no limit had been imposed)	0 - 49,630	0 - 21	21 - 239	252 - 1,072	1,133 - 49,630
5. Percent screenout per search question		100.0 - 99.96	99.96 - 99.60	99.60 - 97.96	97.96 - 10.71

TABLE B-II

SCREENING AND ITERATIVE SEARCH DATA
SUBSTRUCTURE SEARCH DEMONSTRATION
NEW YORK, SEPTEMBER 1966

Question Number	Compounds Passing Screens	Percent Screenout	Number of Answers(1)	Projected Number of Answers(2)
1	201	99.6	0	0
2	194	99.7	0	0
3	225	99.6	0	0
4	206	99.6	6 (63)	20
5	5544	90.0	6 (229)	144
6	6	99.99	0	0
7	1072	98.07	6 (472)	13
8	789	98.58	0	0
9	68	99.88	0	0
10	0	100.0	0	0
11	718	98.71	0	0
12	10	99.99	5	5
13	3	99.99	0	0
14	2	99.99	0	0
15	2483	95.52	6 (1267)	11
16	252	99.6	0	0
17	4216	92.39	0	0
18	1055	98.07	0	0
19	45	99.92	6 (14)	19
20	13	99.99	0	0
21	450	99.19	6 (39)	69
22	963	98.27	0	0
23	77	99.88	6 (9)	51
24	1	99.99	0	0
25	2590	95.33	6 (17)	911
26	138	99.81	6 (26)	25
27	556	99.0	6 (226)	14
28	40	99.93	0	0
29	172	99.79	2	2
30	395	99.29	6 (6)	395
31	1016	98.17	6 (358)	16
32	43	99.92	0	0
33	525	99.06	6 (13)	262
34	9754	82.40	0	0
35	156	99.72	0	0

(1) For questions which had fewer than 6 hits, all compounds passing the screens had to be iteratively searched. For questions which reached the maximum cutoff of 6 hits, the number in parentheses indicates the number of compounds which had to be iteratively searched up to that point.

(2) For questions which had the maximum cutoff of 6 hits, the projected number of hits was calculated as follows:

$$\text{Projected number} = \frac{\text{Number of compounds passing screens}}{\text{Number of iterative searches}} \times 6$$

SCREENING AND ITERATIVE SEARCH DATA (cont'd)

Question Number	Compounds Passing Screens	Percent Screenout	Number (1) of Answers	Projected Number (2) of Answers
36	8	99.99	1	1
37	81	99.88	2	2
38	18	99.97	0	0
39	2685	95.16	1	1
40	156	99.72	2	2
41	367	99.34	0	0
42	409	99.2	6 (95)	26
43	18	99.97	1	1
44	319	99.43	6 (22)	87
45	357	99.4	6 (244)	9
46	32	99.95	0	0
47	3259	94.12	0	0
48	55	99.85	0	0
49	2493	95.5	0	0
50	321	99.4	6 (15)	128
51	2068	96.27	6 (11)	1130
52	36	99.9	1	1
53	34	99.9	4	4
54	1727	97.89	6 (47)	218
55	36	99.9	6 (24)	9
56	315	99.44	0	0
57	493	99.12	6 (33)	89
58	725	98.7	1	1
59	252	99.6	6 (13)	126
60	6	99.99	2	2
61	25	99.96	5	5
62	24,243	56.24	6 (175)	824
63	1217	97.81	6 (71)	101
64	201	99.6	6 (15)	80
65	1824	96.71	6 (10)	1094
66	946	98.3	0	0
67	47	99.9	0	0
68	56	99.85	0	0
69	2	99.99	1	1
70	2043	96.3	5	5
71	9	99.99	0	0
72	1500	97.30	0	0
73	0	100.0	0	0
74	1364	97.52	0	0
75	1637	97.05	0	0
76	79	99.86	1	1
77	849	98.47	0	0
78	3159	94.20	6 (240)	78
79	367	99.4	1	1
80	46	99.9	0	0

SCREENING AND ITERATIVE SEARCH DATA (cont'd)

Question Number	Compounds Passing Screens	Percent Screenout	Number of Answers (1)	Projected Number of Answers (2)
81	26	99.96	0	0
82	17	99.97	0	0
83	13	99.97	0	0
84	21	99.97	6 (6)	21
85	1857	96.6	2	2
86	21	99.96	6 (7)	20
87	0	100.0	0	0
88	0	100.0	0	0
89	475	99.15	6 (6)	475
90	34	99.9	0	0
91	143	99.75	5	5
92	*			
93	1	99.99	0	0
94	1059	98.07	6 (87)	72
95	0	100.0	0	0
96	239	99.6	0	0
97	4596	91.71	6 (26)	1057
98	36	99.9	0	0
99	4117	92.3	6 (1455)	17
101	10	99.99	6 (6)	10
102	0	100.0	0	0
103	24	99.96	6 (6)	24
104	11	99.99	6 (7)	10
105	230	99.6	3	3
106	4	99.99	4	4
107	868	98.45	5	5
108	5963	89.24	6 (364)	95
109	0	100.0	0	0
110	360	99.4	6 (12)	180
111	4579	91.72	6 (814)	32
112	329	99.4	0	0
113	0	100.0	0	0
114	20	99.98	0	0
115	253	99.6	6 (6)	253
116	1294	97.37	5	5
117	32	99.95	0	0
118	425	99.24	6 (319)	8
119	0	100.0	0	0
120	9	99.99	0	0
121	16,716	69.83	6 (1357)	67
122	1426	97.43	6 (128)	66
123	51	99.8	0	0
124	1278	97.4	6 (283)	27
125	324	99.4	0	0

* Bibliography search only was performed on four specific compounds which had been identified by structure and Registry Number.

SCREENING AND ITERATIVE SEARCH DATA (cont'd)

Question Number	Compounds Passing Screens	Percent Screenout	Number of Answers (1)	Projected Number of Answers (2)
126	7	99.99	0	0
127	187	99.7	0	0
128	259	99.6	0	0
129	1463	97.35	0	0
130	2839	94.88	6 (1530)	9
131	333	99.4	5	5
132	2685	95.16	6 (1045)	13
133	517	99.07	0	0
134	0	100.0	0	0
135	2863	99.88	6 (160)	106
136	3256	94.12	6 (1006)	16
137	526	99.08	6 (115)	27
138	596	98.9	6 (6)	596
139	426	99.24	6 (8)	320
140	134	99.76	1	1
141	0	100.0	0	0
142	17	99.98	3	3
143	21	99.96	6 (6)	21
144	81	99.86	0	0
145	41	99.94	0	0
146	49.630	10.41	6 (6297)	45
147	0	100.0	0	0
148	0	100.0	0	0
149	149	99.74	6 (6)	149
150	34	99.95	0	0
151	1133	97.96	1	1
152	795	98.57	6 (14)	340
153	27.896	49.65	6 (226)	725
154	203	99.64	0	0
155	5954	89.24	6 (17)	1985
156	2295	95.87	6 (593)	23
157	493	99.12	0	0
158	336	99.4	6 (90)	22
159	6451	88.36	6 (9)	4296
160	3553	93.59	6 (12)	1777
161	14.919	75.37	6 (145)	612
162	14.360	74.08	6 (1895)	43
163	1852	96.6	0	0
164	260	99.6	0	0
165	274	99.61	0	0
166	1346	97.53	6 (465)	16
167	1543	97.22	6 (69)	133
168	0	100.0	0	0
169	155	99.72	6 (16)	58
170	6	99.99	0	0

SCREENING AND ITERATIVE SEARCH DATA (cont'd)

Question Number	Compounds Passing Screens	Percent Screenout	Number of Answers (1)	Projected Number of Answers (2)
171	857	98.46	6 (62)	84
172	7	99.99	0	0
173	0	100.0	0	0
174	4895	91.17	6 (1690)	15
175	0	100.0	0	0
176	190	99.7	6 (77)	15
177	3068	94.47	6 (127)	144
178	4	99.99	0	0
179	5	99.99	0	0
180	308	99.45	6 (17)	103
181	891	98.4	6 (85)	62
182	10	99.98	6 (6)	10
183	47	99.92	6 (44)	6
184	14,806	73.3	6 (6)	14,806

APPENDIX C

SCREENS

SCREENS

For the Substructure Search demonstration, CAS utilized eight different screens that together contained approximately 1350 screen items, as listed in Table C-I below. A screen item may include not only the identification of a structural feature, but also a numerical indication of the number of times it appears in a structure. For example, one screen item might require one occurrence of the fragment C-C in a compound, while another screen might require two occurrences of the same fragment. Each screen

TABLE C-I

Screen	Number of Screen Items
Atom Counts	44
Ring Counts	20
Element Counts	170
Bond Counts	114
Atom-Bond-Atom "Triplets"	419
First-Level Connectivities ("Moieties")	502
Salt, Ammoniate, and Hydrate Fragments	44
Ring Sizes and Specific Structural Characteristics	31

item is chosen according to chemists' intuition and the results of earlier experience; each selected item is assigned one of the 2000 bit-indicator positions set aside for that purpose.

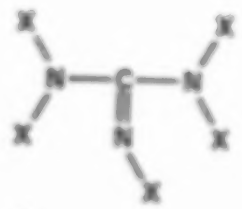
All the screen items taken together constitute a screen dictionary from which appropriate screens are identified for each Substructure Search question. The frequency with which each screen item occurred in the demonstration file of 55,396 compounds was also available to assist in question coding.

Explanation of Screen Types

The following explanations and examples illustrate the screen types used in the demonstration. The "Appropriate Screen Items" specified in each example apply only to the screen type under discussion. In actual searches, appropriate screen items are chosen from several screen types.

1. TOTAL ATOM COUNT SCREEN eliminates all compounds having fewer than a specified number of nonhydrogen atoms.

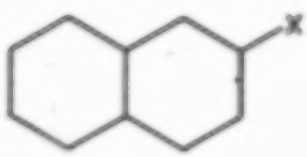
   EXAMPLE:

Question	Appropriate Screen Item
	Require a count of 9 or more atoms for each potential answer.

X = any nonhydrogen atom

2. RING COUNT SCREEN eliminates all compounds having fewer than a specified number of rings. Rings are defined according to the Ring Index rules (i.e., the minimum number of scissions of ring bonds required to produce a completely acyclic structure).

   EXAMPLE:

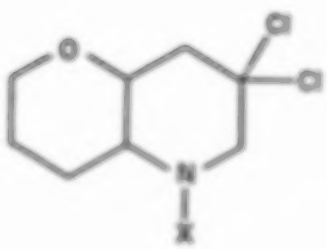
Question	Appropriate Screen Item
	Require a count of 2 or more rings for each potential answer.

X = any nonhydrogen atom

3. ELEMENT COUNT SCREEN eliminates those structures having fewer than a specified number of atoms of a given element or elements.

There is at least one screen item in the Element Composition category for each element of the periodic table. The more commonly occurring elements (e.g. C,N,O,S, and halogens) have additional screen items to allow for higher frequency counts. For example, in addition to the screen item for a single Cl in a structure, there are screen items for two, three, five, seven, and nine or more Cl's in a structure.

EXAMPLE:

Question	Appropriate Screen Items
	Require 2 or more Cl 9 or more C 1 or more N

X = any nonhydrogen atom

4. BOND TYPE AND COUNT SCREEN eliminates all structures having fewer than a specified number of a given type of bond or bonds. Bond types are defined as follows:

Bond Symbol	Bond Significance
1	Single bond, acyclic
2	Double bond, acyclic
4	Triple Bond, acyclic
B	"Don't Care" (any bond is acceptable)
C	Any ring bond
J	Single bond, cyclic
K	Double bond, cyclic

Bond Symbol	Bond Significance
L	Any bond that is part of a fully conjugated system of single and double ring bonds
M	Triple bond, cyclic
W	Any chain bond

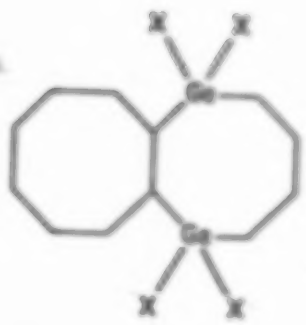
EXAMPLE:

Question	Appropriate Screen Items
	Require 6 or more L bonds 1 or more 1 bonds 2 or more 2 bonds

X = C, N, or O

5. ATOM-BOND-ATOM "TRIPLETS" SCREEN eliminates all structures having fewer than a specified number of "triplets". A triplet is defined by identifying two connected atoms and their connecting bond. Atom-bond-atom triplets are included only for the 12 most populous elements on file: B,Br,C,Cl,F,I,N,O,P,S,Si,Sn. Specific bond types are identified as listed in Screen 4 for most pairs of elements, but for less common elements (i.e., B,I,Si,Sn), many screen items describe only generic-level bonds.

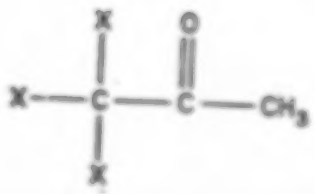
EXAMPLE:

Question	Appropriate Screen Item
	Require 11 or more C-C triplets

X = any atom

6. FIRST-LEVEL CONNECTIVITY SCREEN ("MOIETIES") eliminates all structures having fewer than a specified number of "moieties." Moieties are defined as the number and type of atoms attached to a central atom together with their connecting bonds. Moiety descriptions are included only for the six most common polyvalent elements, namely: C,N,O,P,S,Si. All bond descriptions in this group are specific.

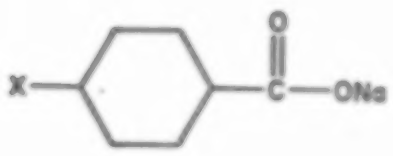
EXAMPLE:

Question	Appropriate Screen Item
	Require at least one C-C-C moiety (C=O)

X = any nonhydrogen atom

7. SALT, AMMONIATE, AND HYDRATE SCREEN eliminates all structural records that do not contain a specified atom or atoms in the "salt portion" of the structural record. Screen items in this category are included only for those elements known to be present in the salt, ammoniate, or hydrate portion of the file. This screen does not include frequency counts.

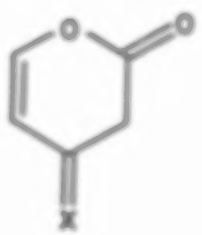
EXAMPLE:

Question	Appropriate Screen Item
	Requires Na to be present in the salt portion of the record of each potential answer.

X = any nonhydrogen atom

8. RING SIZES AND OTHER SPECIFIC STRUCTURAL CHARACTERISTICS. These screens eliminate all structures that do not contain certain special chemical fragments not included in the above screen categories. In this category there are screen items for specific ring sizes of 3 to 19 atoms inclusive plus a screen item for rings containing 20 or more atoms. These ring sizes are applicable to all possible cyclic paths in a structure (e.g., anthracene has rings of 6, 10, and 14 atoms). Other screen items included in this category are various groups such as 6-membered carbocycle, steroid nucleus, etc., and some generic screens such as any metal, any halogen, any hydrocarbon.

EXAMPLE:

Question	Appropriate Screen Item
	Requires a six-membered heterocycle plus other applicable screens for each potential answer.

X = any nonhydrogen atom

APPENDIX D

DEMONSTRATION FILE CHARACTERISTICS

TABLE D-I

NUMBER OF OCCURRENCES OF VARIOUS ELEMENTS*

Elem.	No. of Occurrences	Elem.	No. of Occurrences	Elem.	No. of Occurrences
Ag	7	Fe	10	Po	6
Al	69	Ga	26	Pt	6
As	326	Ge	221	Ru	1
Au	1	Hg	319	S	20,803
B	819	I	732	Sb	99
Ba	1	K	8	Se	247
Be	1	La	2	Si	1,390
Bi	17	Li	34	Sn	534
Br	2,639	Mg	35	Sr	1
C	-	Mn	2	T	38
Ca	9	Mo	3	Ta	2
Cd	1	N	>55,000	Te	42
Cl	8,701	Na	15	Th	1
Co	3	Nb	1	Ti	18
Cr	20	Ni	3	Tl	17
Cu	18	O	>55,000	U	1
D	256	P	5,548	V	44
Eu	1	Pb	98	Zn	55
F	7,438	Pd	2	Zr	9

*Redundancy exists in this Table since the figures are based on the number of occurrences of compounds containing a specific number and type of atom-bond-atom combinations, or "triplets". For example, methanesulfonic acid contains one C——S bond, two S=O bonds and one S——O bond; this accounts for three occurrences of sulfur in the Table.

TABLE D-II

NUMBER OF COMPOUNDS CONTAINING VARIOUS ELEMENTS
AS SALT, AMMONIATE, OR HYDRATE FRAGMENTS

Metal Salts				Non-Metal Salts*		Other	
Elem.	No. of Compds.	Elem.	No. of Compds.	Elem.	No. of Compds.	Elem.	No. of Compds.
Ag	21	K	113	Br	595	$B(BH_3)$	13
Al	9	Li	18	Cl	3108	$N(NH_3)$	63
Au	14	Mg	12	F	16	$O(H_2O)$	15
Ba	23	Mn	10	I	507		
Be	3	Na	445				
Bi	1	Nd	1				
Ca	45	Ni	3				
Cd	2	Pb	8				
Ce	1	Pd	1				
Co	10	Pr	1				
Cs	8	Pt	1				
Cu	25	Rb	2				
Dy	1	Sb	1				
Eu	1	Sn	3				
Fe	9	Sr	2				
Ga	1	V	1				
Hg	9	Zn	26				
Ho	1	Zr	2				

*Includes only single-atom anions

TABLE D-III

ELEMENTS NOT APPEARING IN ANY COMPOUNDS OF THE DEMONSTRATION FILE

Ac	Fr	Ne	Rn
Am	Gd	No	Sc
Ar	He	Np	Sm
At	Hf	Os	Tb
Bk	In	Pa	Tc
Cf	Ir	Pm	Tm
Cm	Kr	Pu	W
Er	Ln	Ra	Xe
Es	Lw	Re	Y
Fm	Md	Rh	Yb

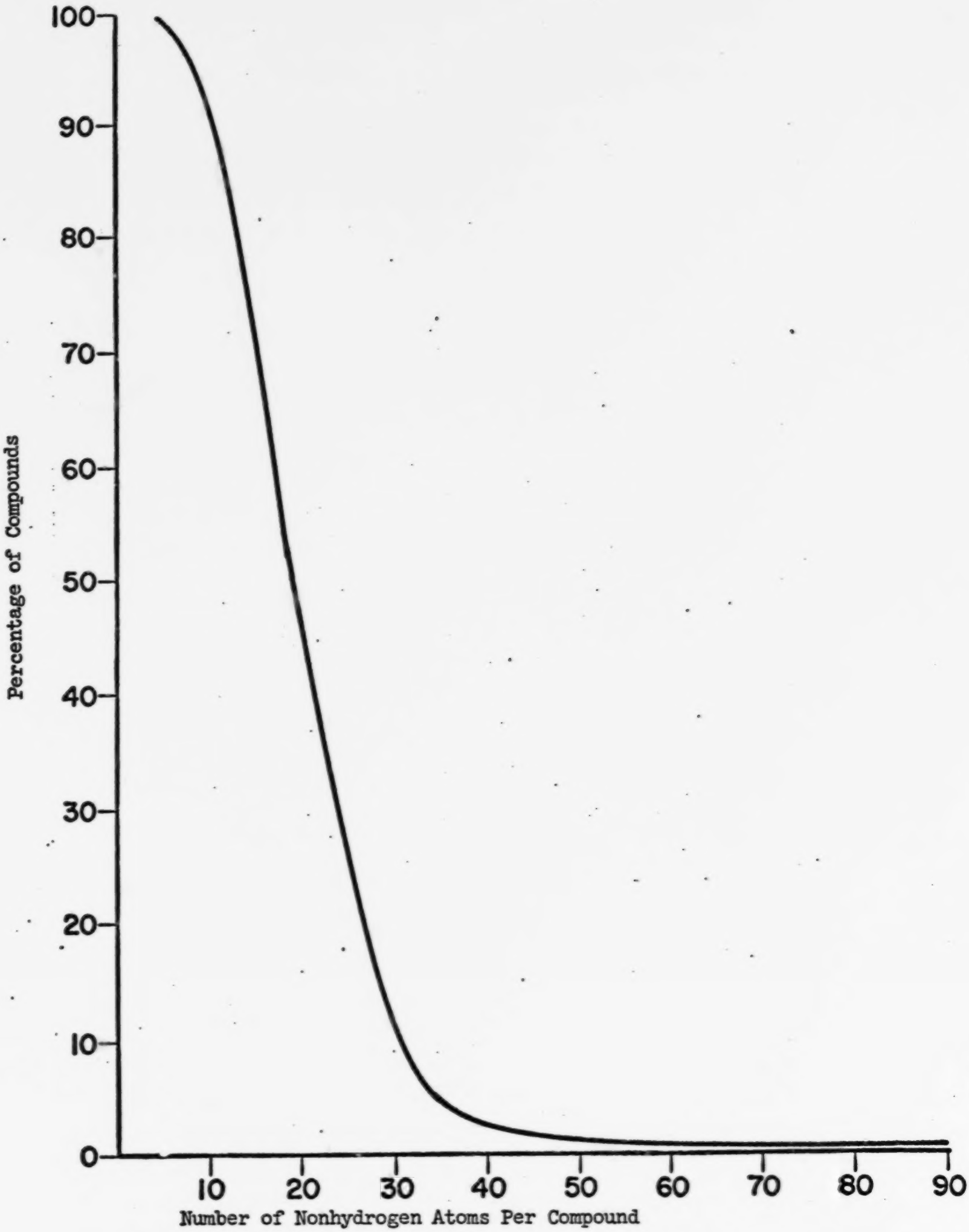


Figure 1 - Distribution of Compounds Containing Different Numbers of Nonhydrogen Atoms

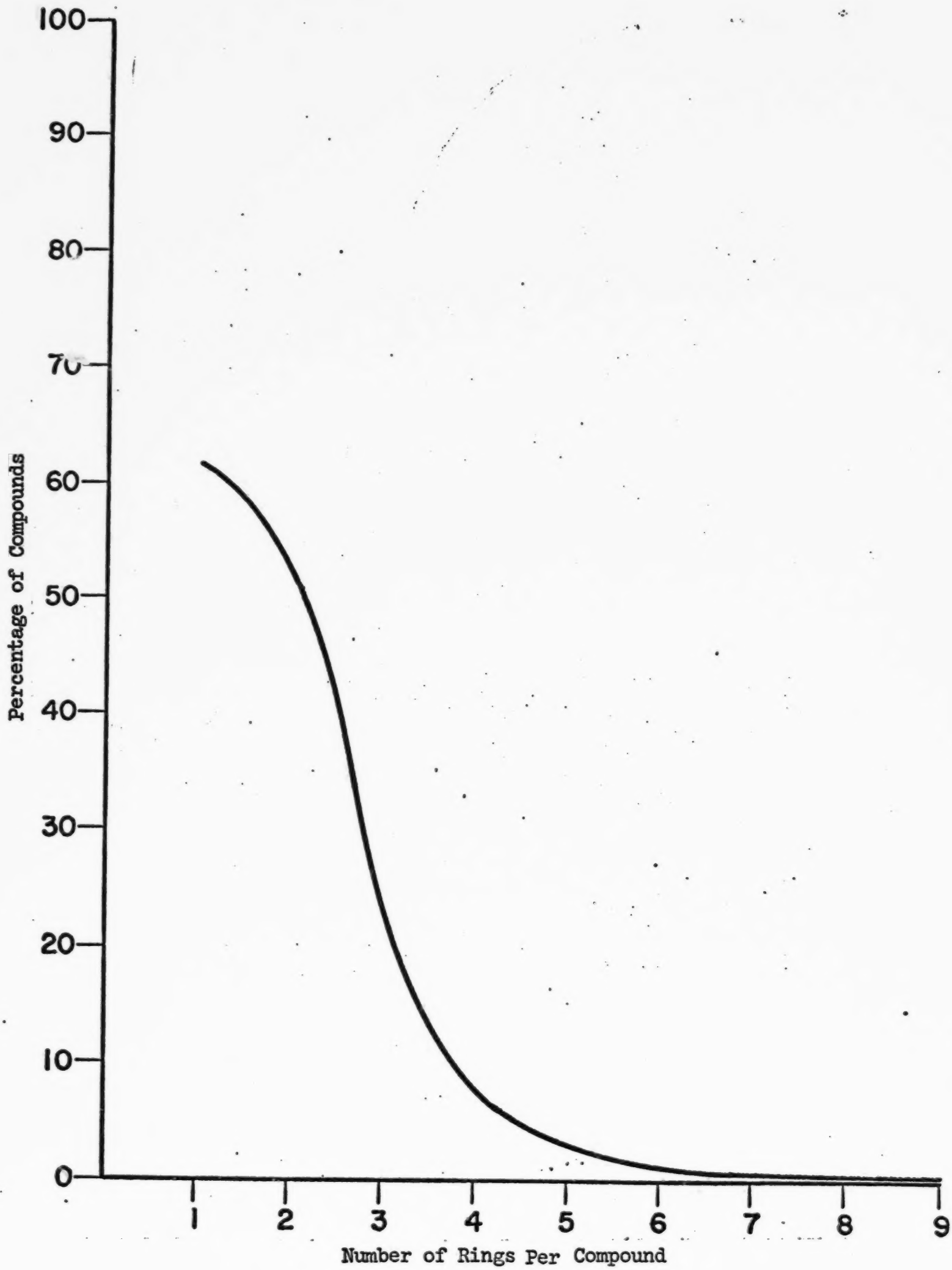


Figure 2 - Distribution of Compounds Containing Different Numbers of Rings

TABLE D-IV

NUMBER OF COMPOUNDS CONTAINING VARIOUS TYPES OF COVALENTLY BONDED ATOM-PAIRS ("TRIPLETS") IN THE DEMONSTRATION FILE*

Atom-pair	Bond Type	No. of Compounds	Atom-pair	Bond Type	No. of Compounds	Atom-pair	Bond Type	No. of Compounds
Br-C	1**	2430	C-N	4	1796	Cl-Si	1	160
Br-N	1	9	C-O	J	6859	F-I	1	1
Br-O	1	1	C-O	K	2	F-O	1	1
Br-O	2	1	C-O	L	76	F-P	1	196
Br-P	1	4	C-O	1	29723	F-S	1	148
Br-S	1	1	C-O	2	28241	F-Si	1	31
Br-Si	1	6	C-O	4	1	I-I	1	8
C-C	J	26396	C-P	J	54	I-O	J	20
C-C	K	10820	C-P	K	1	I-O	j	23
C-C	L	35594	C-P	1	1121	I-O	2	2
C-C	M	10	C-P	2	79	I-P	1	1
C-C	1	47822	C-S	J	3246	I-S	1	1
C-C	2	7064	C-S	K	1	I-Si	1	1
C-C	4	885	C-S	L	5	N-N	J	1660
C-I	J	22	C-S	1	6464	N-N	K	258
C-I	1	595	C-S	2	1247	N-N	L	170
C-N	J	13490	Cl-I	1	3	N-N	1	2623
C-N	K	4530	Cl-N	1	53	N-N	2	1092
C-N	L	4858	Cl-O	1	246	N-N	4	201
C-N	M	1	Cl-O	2	249	N-O	J	352
C-N	1	27396	Cl-P	1	127	N-O	1	972
C-N	2	3945	Cl-S	1	115	N-O	2	5332

*Only atom pairs in which any of the 10 most common elements (i.e., Br, C, Cl, F, I, N, O, P, S, Si) is bonded to another are included in this table.
**See Appendix C, Screen 4 for identification of bond types.

(continued)

Atom-pair	Bond Type	No. of Compounds	Atom-pair	Bond Type	No. of Compounds	Atom-pair	Bond Type	No. of Compounds
N-P	J	17	O-O	1	111	P-P	4	
N-P	L	27	O-P	J	71	P-S	J	
N-P	1	314	O-P	1	1530	P-S	1	277
N-P	2	64	O-P	2	1239	P-S	2	400
N-P	4	4	O-S	J	56	S-S	J	55
N-S	J	392	O-S	1	1645	S-S	1	327
N-S	K	2	O-S	2	4329	S-Si	1	2
N-S	1	1886	O-Si	J	67	S-Si	2	1
N-S	2	56	O-Si	1	264	Si-Si	J	2
N-Si	J	8	O-Si	2	2	Si-Si	1	31
N-Si	1	37	P-P	J	2			
O-O	J	13	P-P	1	5			

APPENDIX E

LIST OF QUESTIONERS AND THEIR AFFILIATIONS

APPENDIX E

LIST OF QUESTIONERS AND THEIR AFFILIATIONS
REMOTE SUBSTRUCTURE SEARCH DEMONSTRATION
NEW YORK, SEPTEMBER 1966

Name	Affiliation
Aaland, Mrs. Sharon	Abbott Lab. North Chicago, Ill. 60064
Aszalos, Dr. A.	Squibb Inst. New-Brunswick, N. J.
Babad, Dr. Harry	Univ. of Denver Dept. of Chem.
Barton, T. J.	Univ. of Florida Dept. of Chemistry Gainesville, Fla. 32601
Bauman, Robert	Colgate-Palmolive Co. 909 River Road Piscataway, N. J. 08854
Benson, Dr. F. R.	Atlas Chem. Ind. Wilmington 99, Delaware Manager, Information Section
Berezin, Dr. G. H.	DuPont Company Explosives Dept. Experimental Station Lab. Wilmington, Delaware
Berger, Dr.	Baxter Labs. Morton Grove, Ill.
Bernier, Dr. Charles L.	The Squibb Inst. for Medical Res. Georges Road New Brunswick, N. J. 08903
Bonanno, S. R.	The Squibb Inst. for Medical Res. Georges Road New Brunswick, N. J. 08903
Bose, Dr. A. K.	Dept. of Chemistry Stevens Institute of Technology Hoboken, N. J. 07030

Name	Affiliation
Boyack, Dr. G. A.	The Upjohn Company Kalamazoo, Michigan
Braswell, Dr. E. H.	Univ. of Conn. Storrs, Conn.
Bristol, D. W.	Chem. Dept. Syracuse Univ. Syracuse, N. Y. 13210
Brown, Horace D.	Merck and Co., Inc Rahway, N. J.
Burt, Dr. G. D.	Harshaw Chemical Co. Cleveland, Ohio 44106
Byck, Joseph S.	Box 408 Havemeyer Columbia University New York, N. Y. 10027
Cardeilhac, Dr. P. T.	Dept. of Physiology and Pharm. Oklahoma State Univ. Stillwater, Oklahoma
Casey, J. P.	Univ. of Virginia Dept. of Chemistry Charlottesville, Va. 22903
Chakrin, A. L.	Univ. of Chicago
Chisolm, R. A.	3M W. Bldg. 201-25 St. Paul, Minn. 55101
Cinnamon, J. M.	Shulton
Clarke, Dr. Donald D.	Fordham Univ.
Crawford, Thomas H.	Dept. of Chemistry Univ. of Louisville Louisville, Ky. 40208
Culvenor, C. C. J.	CSIRO, Australia
DeStephen, Tony	Harshaw Chem. Co. Cleveland, Ohio 44106
Donovan, Miss Kathryn M.	Pennsalt Chemicals Corp. 900 First Ave. King of Prussia, Pa. 19406
Drew, Dr. Howard F.	Proctor and Gamble Research Division Miami Valley Labs.

Name	Affiliation
DuDock, Dr. B. S.	Dept. of Biochemistry Cornell Univ. Ithaca, N. Y.
Dutton, Herbert	Northern Regional Research Lab. 1815 N. University Peoria, Ill.
Ebert, Miss Helen M.	Smith, Kline and French
Eddy, Dr. L. P.	Western Washington State College
Elston, Dr. C. T.	DuPont of Canada Research Center Kingston, Ontario
Fallon, Dr. Frances	The Wm. S. Merrell Co. Cincinnati, Ohio 45215
Fetterolf, Dr. L. M.	Smith, Kline and French 1500 Spring Garden Street Philadelphia, Pa.
Finkbeiner, Dr.	General Electric Res. Box 8 Schenectady, N. Y.
Foote, Dr. H. E.	Avi Publ. Co.
Fraction, George	Eli Lilly and Co. Indianapolis, Ind. 46205
Franck, Dr. Richard W.	Chemistry Dept. Fordham Univ. Bronx, N. Y. 10458
Frank, Dr. S.	American Cyanamid Co. Central Research Div. Stamford, Conn.
Friedman, Dr. Herbert A.	Sloan-Kettering Institute 145 Boston Post Road Rye, N. Y. 10580
Gans, Richard	Frick Chem. Lab. Princeton Univ. Princeton, N. J. 8540
Garwig, Paul L.	F.M.C. Corp Box 8 Princeton, N. J.

Name	Affiliation
Gassmann, Dr. Paul	Chem. Dept. Ohio State University
Gelberg, Alan	Diamond Alkali Company
Gerson, H.	Allied Chemical Corp. Box 14 Hawthorne, N. J.
Giddings, W. P.	Pacific Lutheran Univ. Tacoma, Washington
Giner-Sorolla, Dr. A.	Sloan-Kettering Inst. 410 E. 68th Street New York
Goldstein, Edward J.	Colgate-Palmolive 909 River Road Piscataway, N. J.
Gosink, T. A.	Old Dominion College Norfolk, Va. 23508
Gough, Dr. S. T. D.	Mobil Chem. Co. Metuchen, N. J.
Gould, Dr. David	Colgate-Palmolive Center Piscataway, N. J.
Grindahl, G. A.	Dow Corning Corp. Midland, Michigan
Gruen, H.	Binghampton, N. Y.
Gudmunsen, Dr. C. H.	Wyeth Labs. Div. Radnor, Pa. 19101
Guiduci, Dr. M. A.	E. R. Squibb New Brunswick, N. J.
Gunther, Dr. W. H. H.	Yale Univ. 333 Cedar Street New Haven, Conn
Haarstad, Dr. V. B.	Tulane Univ. New Orleans, La.
Haggard, Dr. R. A.	Rohm and Haas Co. Springhouse, Penn. 19477
Hall, Dr. H. J.	Esso Research P. O. Box 51 Linden, N. J.

Name	Affiliation
Hamaker, Dr. J. W.	Dow Chem. Walnut Creek, California
Hayward, H. W.	U. S. Patent Office R and D 1406 G. Street Washington, D. C.
Heckman, Robert A.	R. J. Reynolds Tobacco Co. Research Dept. Winston-Salem, N. C.
Heidt, Dr. L. J.	M.I.T. Cambridge, Mass.
Hollinden, S.	Eli Lilly and Co. McCarty and Alabama Streets Indianapolis, Indiana
Holly, Lloyd A.	Industry Liaison Office Research Labs. Edgewood Arsenal, Md.
Hopps, Dr. Harvey	Aldrich Chemical Co. 2369 N. 29th Street Milwaukee, Wisconsin 53210
Iorio, E. James	Chemistry Dept. Northeastern Univ. Boston, Mass.
Jacobs, Dr. R. L.	Maume Chem. Co. 1310 Expressway Drive Toledo, Ohio
Kaback, Dr. S. M.	Esso Research and Eng. Linden, N. J.
Kanter, M. J.	Dept. of Chemistry Univ. of Ill.
Kassel, R. J.	Edgewood Arsenal Chem. Research Labs. Md.
Kazama, Yoshiteru	Stevens Inst. of Tech. P. O. Box 1236 Castle Point Jtn. Hoboken, N. J. 07030
Kellett, Dr. J. C.	N.S.F. 1800 K. Street, N. W. Washington, D. C.

Name	Affiliation
Kerber, Dr. Robert C.	Dept. of Chemistry State Univ. of New York Stony Brook, N. Y. 11790
Korman, J.	Upjohn Co. Kalamazoo, Michigan
Kriman, Dr. M. M.	Allied Chem. Corp. Morristown, N. J.
Kuntz, I.	Enjay Polymer Labs. P. O. 45 Linden, N. J. 07036
Kurtz, Arthur Peter	Box 408 Havemeyer Hall Dept. of Chemistry Columbia University New York, New York 10027
Kwiatek, Dr. J.	U. S. Industrial Chemicals Co. 1275 Section Road Cincinnati, Ohio 45237
LaMontagne, M. P.	Duquesne University Dept. of Chemistry Pittsburgh, Pa. 15219
Landers, J. O.	Dept. of Chem. Ohio State Univ. Columbus, Ohio 43210
Langer, Dr. S. H.	Chem. Engr. Dept. Univ. of Wisconsin Madison, Wisconsin
Levine, Dr. R.	Univ. Pittsburgh Chemistry Department Pittsburgh, Pa. 15213
Libby, Louis H.	Research Triangle Park North Carolina Science and Technology Research Center North Carolina
Liebman, J. F.	Brooklyn College (Mail to: 2962 Brighton 8th Street Brooklyn, N. Y.
Lipowitz, Dr. J.	Dow Corning Corp. Midland, Michigan (Phys. Chem. Res. Dept.)

Name	Affiliation
Liu, Mr. Joseph Ko-Chiung	Dept. of Chemistry McGill University Montreal 2, P.Q., Canada
Long, Gary J.	Dept. of Chemistry Syracuse Univ. Syracuse, N. Y. 13210
Longenecker, W. H.	Fort Detrick Fred., Md. 21701
Lyle, Dr. R. E.	Dept. of Chem. Univ. of New Hampshire Durham, N. H.
Malkiewich, E. J.	Hoffmann-LaRoche Nutley, N. J. 07110
Maizell, Dr. R. E.	Olin Mathieson Chemical Corp.
Marsh, Dr. John L.	Ciba Pharmaceutical Co. Morris Ave. Summit, N. J.
Marshall, Dr. W. J.	DuPont Pigments Dept. 256 Vanderpool Street Newark, N. J.
Matthews, Fred W.	Canadian Industried Ltd. McMasterville, Quebec
McCarthy, Miss J.	Monsanto Co. 1700 South 2nd Street St. Louis, Mo. 63177
McKelvie, Prof. Neil	Dept. of Chemistry City College (city U. of N. Y.) Convent Ave. and 140 Street N. Y. 10031
Milewich, Dr. L.	Johns Hopkins University School of Medicine Baltimore, Md.
Mitchell, Leonard D.	Herner and Co. Washington, D. C.
Montague, Miss B. A.	DuPont Wilmington, Delaware
Narvaeg, Dr. R.	DuPont Company Experimental Station Wilmington, Delaware

Name	Affiliation
Notation, Dr. A. D.	Univ. of Minnesota Biochemistry Dept.
Nutting, N. H.	University of California
Odstrchel, Dr. G.	Duquesne Univ.
Orchin, Dr. M.	Univ. of Cincinnati Dept. of Chem. Cincinnati, Ohio
Parker, William L.	Dow Chemical Co. P. O. Box 400 Qayland, Mass.
Pathak, Balai Chand	School of Pharmacy Dept. of Medicinal Chemistry University of Buffalo N. Y. 14214
Phillips, Dr. A. P.	Burroughs Wellcome Co.
Pinkus, Dr. J. L.	Univ. of Pittsburgh Dept. of Chemistry Pittsburgh, Pa. 15213
Regan, Dr.	Baxter Labs. Morton Grove, Ill.
Rice, Dr. Charles	Eli Lilly and Co. Indianapolis, Ind. 46205
Roberts, Dr. D. L.	R.J. Reynolds Tobacco Co. Winston-Salem North Carolina 27101
Ross, Joseph	Indiana University South Bend, Indiana 46615
Santoro, Angelo	Hunter College Park Ave. and 69th Street N. Y.
Schafter, Dr. C. D.	Inst. Für Documentation 6 Frankfurt/Main 1 Vogtstr. 50
Scheffler, Dietmar	Univ. of Delaware Newark, Delaware
Schlessinger, Dr. G. G.	Newark College of Engineering Chemistry Dept. 323 High Street Newark, N. J. 07102
Schramm, William	Food and Drug Administration Washington, D. C.

Name	Affiliation
Scott, P. M.	Food and Drug Directorate Ottawa, Canada
Shwayder, W. M.	Shwayder Chemical Metallurgy Corp. 684 E. Woodbridge Detroit 26, Michigan
Simmons, Dr. Noel	State Univ. College Elmwood Avenue Buffalo, New York 14222
Skoza, Lorant	1856 85th Street New York, N. Y. 10028
Slater, J.	Manager R and D Southern Nitrogen Co. Savannah, Georgia
Slavin, Donald	Sadtler Research Labs. 3316 Spring Garden Street Philadelphia, Pa. 19104
Smith, James H.	Univ. of California
Srinivasan, Dr. V. R.	L.S.U. Baton Rouge, Louisiana
Stanfield, Dr. M. K.	Dept. of Biochem. Tulane Med. School 1430 Tulane New Orleans, La 70112
Starkey, R. J.	Perry Rubber Co. 1875 Harsh Ave. S. E. Massilon, Ohio
Stern, Dr. R. L.	Northeastern Univ. Dept. of Chem. Boston, Mass.
Stolow, Dr. R. D.	Tufts Univ. Chem. Dept. Medford, Mass. 02155
Stucky, Galen	Chem. Dept. Univ. of Illinois
Swartz, J.	Olin Research Center Tech. Information Serv. New Haven, Conn.
Theilheimer, Dr. William	Hoffmann-LaRoche Nutley, N. J. 07110

Name	Affiliation
Thirtle, Dr. J. R.	Eastman Kodak Corp.
Thompson, Dr. M. W.	Rutgers University
Tillmanns, Dr. Emma June	Atlas Chemical Industries Wilmington, Delaware 19899
Triner, W. J.	General Aniline and Film Corp.
Usher, Dr. D. A.	Cornell University Ithaca, New York 14850 (Baker Laboratory)
Van Cot, Dr. J. G.	DuPont Co. Wilmington, Delaware
Viola, A.	Northeastern Univ. Boston, Mass.
Voo, D.	C. F. Braun and Co. Murray Hill, N. J.
Waring, Sister Mary Grace	Marymount College Salina, Kansas
Weakley, M. L.	Nipak Pryor, Oklahoma
Wei, P. H. L.	Wyeth Labs., Inc. Radnor, Pa.
Wilcox, Dr. C. F.	Chemistry Dept. Cornell Univ. Ithaca, N. Y.
Williamson, K.	Dept. of Chem. Mount Holyoke College South Hadley, Mass. 01075
Yaktin, H. K.	Hess and Clark Research Farm Ashland, Ohio 44805
Youker, John	Rensselaer Polytechnic Inst. Troy, N. Y. Dept. of Chemistry
Young, Prof. J. A.	Kings College Wilkes-Barre, Pa.

Name	Affiliation
Young, Dr. Lewis	Dow Chem.
Zabik, Matthew J.	Dept. of Entomology Michigan State Univ. East Lansing, Michigan 48823
Zwick, Dr. M. M.	American Cyanamid 1937 W. Main Street Stamford, Conn.

APPENDIX F

EXAMPLES OF QUESTIONS AND RETRIEVED ANSWERS

EXAMPLES OF QUESTIONS AND RETRIEVED ANSWERS

This appendix contains examples of five questions presented at the New York demonstration. Each is followed by a reproduction of the answers retrieved by substructure search. In each case, the answers are composed of the Registry Number of the compound containing the substructure, its molecular formula, Preferred CA Name, bibliographic citations, and a graphic representation of the structure.

The answer sheets are self-explanatory with the possible exception of the bibliographic citations. Therefore, a brief explanation follows as to their interpretation.

Chemical Abstracts

An example of a CA reference is: c63:p13271d. The "c" designates that this is a CA reference, "63" is the volume number, and "p" signifies that the reference is an abstract of a patent. The omission of the "p" indicates that the citation is other than a patent. The numerals that follow designate the column number. The concluding "d" in this example represents the portion of Page 13271 on which the abstract is located.

SOCMA

References from the Synthetic Organic Chemical Manufacturers Association handbook start with the abbreviation "socma" followed by a dash, followed by the page number and a letter designating the position of the reference on the page (e.g., socma-657b).

Merck

References from the Merck Index of Chemicals and Drugs start with the abbreviation "merck" followed by a dash, followed by a page number and a letter designating the side of the two-column page in which the reference appears (e.g., merck-0777r).

CZ

References from Chemishes Zentralblatt start with the abbreviation "cz" followed by a dash, followed by the page number (e.g., cz-5655).

SUBSTRUCTURE SEARCH DEMONSTRATION
Chemical Abstracts Service
REQUEST FORM

Name: Example 1 ***Batch Number:***

Affiliation: ***Question Number:***

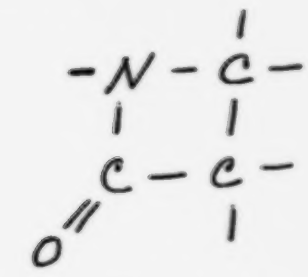
Date:

Chem:

Syst:

Delivery:
☐ ***Pickup*** ☐ ***Mail***

Substructure Request:



β-lactams Prepare to accept large no. of answers
Related to penicillin & cephalosporin
Writing monograph on subject.

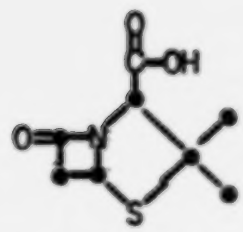
EXAMPLE 1 - ANSWER 1

REGISTRY NO. = 87,536

$C_8H_{11}NO_3S$

Preferred Name: 4-Thia-1-azabicyclo<3.2.0>heptane-2-carboxylic acid, 3,3-dimethyl-7-oxo-

socma-657b
ca2:44628
cz-5655
merck-0777r
ca3:p13271d



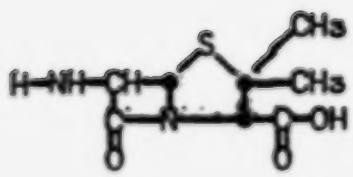
EXAMPLE 1 - ANSWER 2

REGISTRY NO. = 1,203,856

$C_8H_{12}N_2O_3S.Na$

Preferred Name: 4-Thia-1-azabicyclo<3.2.0>heptane-2-carboxylic acid, 6-amino-3,3-dimethyl-7-oxo, sodium salt

ca2:130169
ca3:18861f



.Na salt

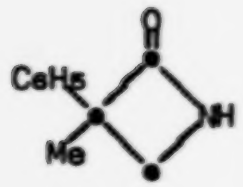
EXAMPLE 1 - ANSWER 3

REGISTRY NO. = 1,623,649

$C_{10}H_{11}NO$

Preferred Name: 2-azetidinone, 3-methyl-3-phenyl-

c63:p042603



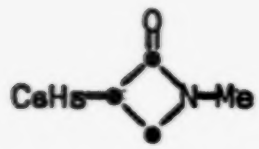
EXAMPLE 1 - ANSWER 4

REGISTRY NO. = 1,623,694

$C_{10}H_{11}NO$

Preferred Name: 2-azetidinone, 1-methyl-3-phenyl-

c63:p042603



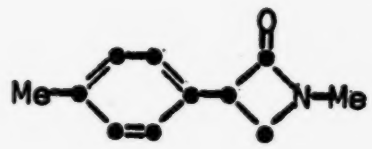
EXAMPLE 1 - ANSWER 5

REGISTRY NO. = 1,623,729

$C_{11}H_{13}NO$

Preferred Name: 2-Azetidinone, 1-methyl-3-p-tolyl-

C63:P04260A



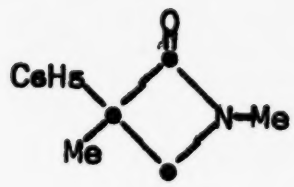
EXAMPLE 1 - ANSWER 6

REGISTRY NO. = 1,746,061

$C_{11}H_{13}NO$

Preferred Name: 2-Azetidinone, 1,3-dimethyl-3-phenyl-

C63:P04260A



SUBSTRUCTURE SEARCH DEMONSTRATION
Chemical Abstracts Service
REQUEST FORM

Name: Example 2

Affiliation: _____

_____

_____

Batch Number: _____

Question Number: _____

Date: _____

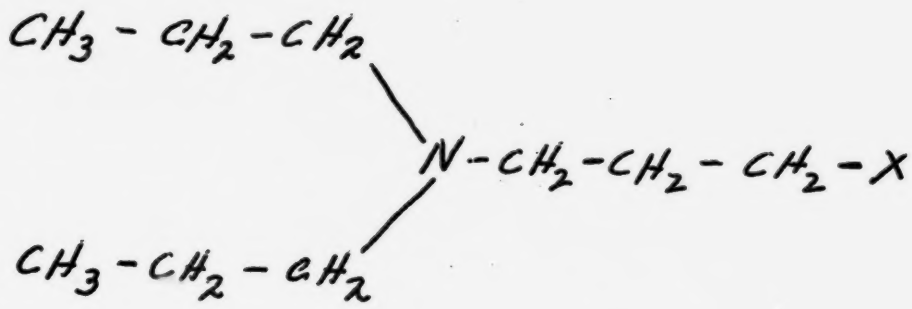
Chem: _____

Syst: _____

Delivery:

☐ ***Pickup*** ☐ ***Mail***

Substructure Request:



X = anything except H
N may have any valence

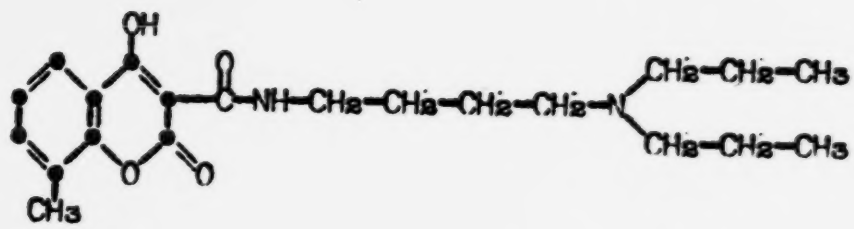
EXAMPLE 2 - ANSWER 1

REGISTRY NO. = 1,556,510

$C_{21}H_{30}N_2O_4$

Preferred Name: Coumarin, 3-<<4-(dipropylamino)butyl>carbamoyl>-4-hydroxy-8-methyl-

CAS:posseod



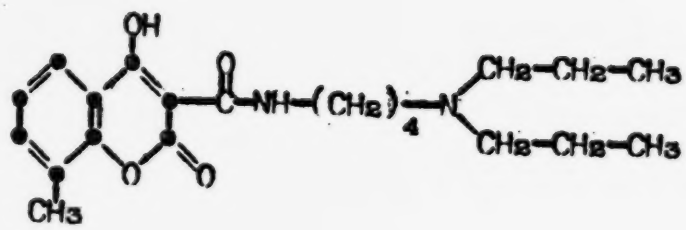
EXAMPLE 2 - ANSWER 2

REGISTRY NO. = 1,556,521

$C_{21}H_{30}N_2O_4.HCl$

Preferred Name: Coumarin, 3<<4-(dipropylamino)butyl>carbamoyl>-4-hydroxy-8-methyl-, hydrochloride

CAS:posseod



.HCl

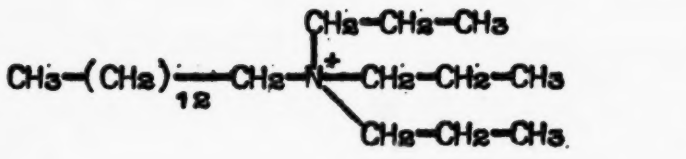
EXAMPLE 2 - ANSWER 3

REGISTRY NO. = 1,112,681

$C_{23}H_{50}N.Br$

Preferred Name: Ammonium, tripropyltetradecyl-, bromide

CAS:022688



$.Br^-$

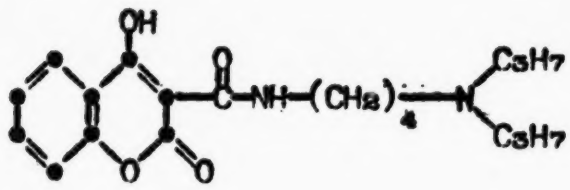
EXAMPLE 2 - ANSWER 4

REGISTRY NO. = 1,435,387

$C_{20}H_{28}N_2O_4$

Preferred Name: Coumarin, 3-<<4-(dipropylamino)butyl>carbamoyl>-4-hydroxy-

CAS:poo531d



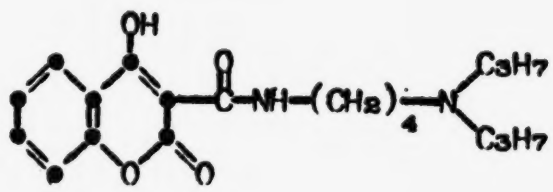
EXAMPLE 2 - ANSWER 5

REGISTRY NO. = 1,435,398

$C_{20}H_{28}N_2O_4.HCl$

Preferred Name: Coumarin, 3-<<4-(dipropylamino)butyl>carbamoyl>-4-hydroxy-, hydrochloride

CSS:poo531d



.HCl

SUBSTRUCTURE SEARCH DEMONSTRATION
Chemical Abstracts Service
REQUEST FORM

Name: Example 3

Affiliation: _____

_____

_____

Batch Number: _____

Question Number: _____

Date: _____

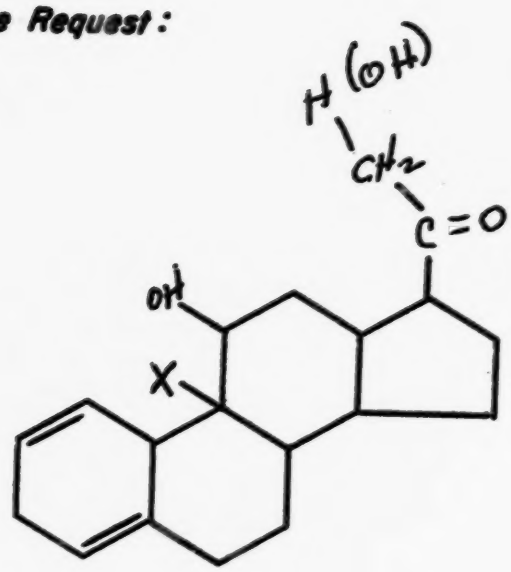
Chem: _____

Syst: _____

Delivery:

☐ ***Pickup*** ☐ ***Mail***

Substructure Request:



X = not Cl, Br, F, I

at least 1 halogen on rings

Form 65-162 8/31/66

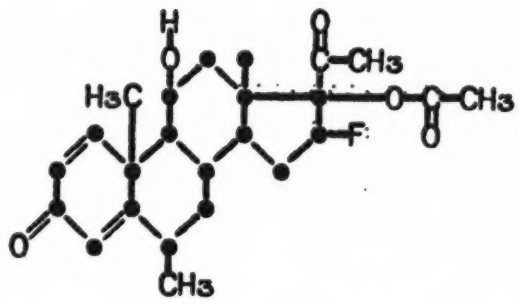
EXAMPLE 3 - ANSWER 1

REGISTRY NO. = 1,597,768

$C_{24}H_{31}FO_5$

Preferred Name: Pregna-1,4-diene-3,20-dione, 16a-fluoro-11b,17-dihydroxy-6a-methyl-, 17-acetate

C57:p919b



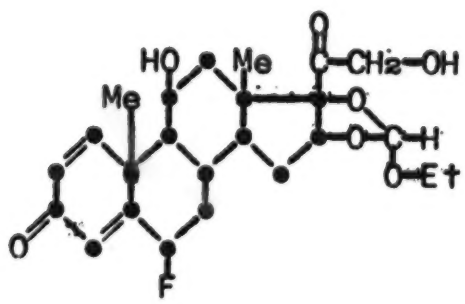
EXAMPLE 3 - ANSWER 2

REGISTRY NO. = 1,683,110

$C_{24}H_{31}FO_7$

Preferred Name: Pregna-1,4-diene-3,20-dione, 6a-fluoro-11b,16a,17,21-tetrahydroxy-, cyclic 16,17-(Et orthoformate)

C50:p3070g



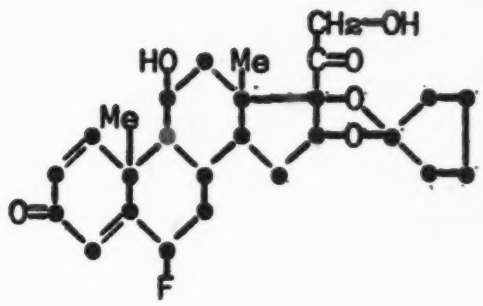
EXAMPLE 3 - ANSWER 3

REGISTRY NO. = 1,735,213

$C_{28}H_{33}FO_6$

Preferred Name: Pregna-1,4-diene-3,20-dione, 6β-fluoro-11b,16a,17,21-tetrah=
ydroxy-, cyclic 16,17-acetal with cyclopentanone

c6o:p3070g



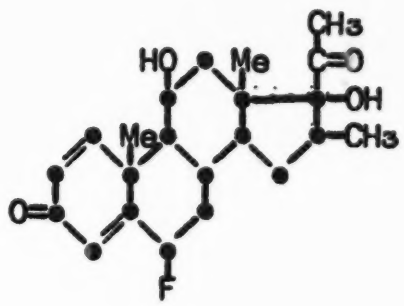
EXAMPLE 3 - ANSWER 4

REGISTRY NO. = 1,800,266

$C_{22}H_{29}FO_4$

Preferred Name: Pregna-1,4-diene-3,20-dione, 6α-fluoro-11b,17-dihydroxy-16α-
-methyl-

c6o:p622f



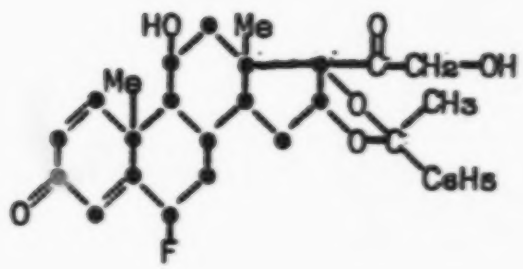
EXAMPLE 3 - ANSWER 5

REGISTRY NO. = 1,841,232

C28H33FO6

Preferred Name: Pregna-1,4-diene-3,20-dione, 6a-fluoro-11b,16a,17,21-tetrah= ydroxy-, cyclic 16,17-acetal with acetophenone

CSO:P3070g



SUBSTRUCTURE SEARCH DEMONSTRATION
Chemical Abstracts Service
REQUEST FORM

Name: Example 4

Affiliation: _____

_____

_____

Batch Number: _____

Question Number: _____

Date: _____

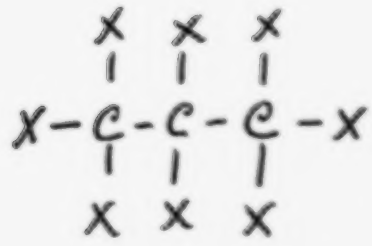
Chem: _____

Syst: _____

Delivery:

☐ *Pickup* ☐ *Mail*

Substructure Request:



X = H, Br, Cl, or F

Atom ct

H	1-3
F	4-6
Cl	1-3
Br	0-1

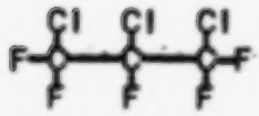
EXAMPLE 4 - ANSWER 1

REGISTRY NO. = 78,175

$C_3Cl_3F_5$

Preferred Name: Propane,1,2,3-trichloropentafluro-

SOCMA-192g
CS2:1049d
CS2:1165b



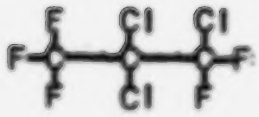
EXAMPLE 4 - ANSWER 2

REGISTRY NO. = 1,599,413

$C_3Cl_3F_5$

Preferred Name: Propane, 1,2,2-trichloropentafluro-

CS1:p1245C
CS2:1116Sb
CS3:28928



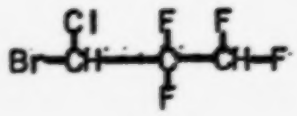
EXAMPLE 4 - ANSWER 3

REGISTRY NO. = 1,645,789

$C_3H_2BrClF_4$

Preferred Name: Propane, 1-bromo-1-chloro-2,2,3,3-tetrafluro-

CSO:P13140h



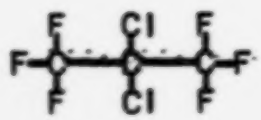
EXAMPLE 4 - ANSWER 4

REGISTRY NO. = 1,652,808

$C_3Cl_2F_6$

Preferred Name: Propane, 2,2-dichlorohexafluoro-

CSS:6679h
CS3:5515C



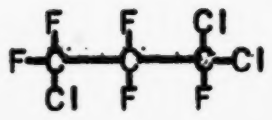
EXAMPLE 4 - ANSWER 5

REGISTRY NO. = 1,652,819

$C_3Cl_3F_5$

Preferred Name: Propane, 1,1,3-trichloropentafluoro-

c53:p14001f
c80:p131408
c82:11165b



SUBSTRUCTURE SEARCH DEMONSTRATION
Chemical Abstracts Service
REQUEST FORM

Name: Example 5

Affiliation: _____

_____

_____

Batch Number: _____

Question Number: _____

Date: _____

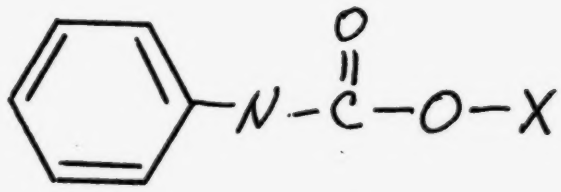
Chem: _____

Syst: _____

Delivery:

☐ ***Pickup*** ☐ ***Mail***

Substructure Request:



Form 65-162 8/31/66

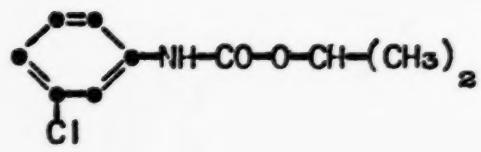
EXAMPLE 5 - ANSWER 1

REGISTRY NO. = 101,213

$C_{10}H_{12}ClNO_2$

Preferred Name: Carbanilic acid, m-chloro-, isopropyl ester

SOCMA-253e
C62:1024e
C62:2188d
C62:p3337b
C62:4520b
C62:5818f
C63:4879C
C63:4880C
C63:4882g
C63:4883C
C63:6256f



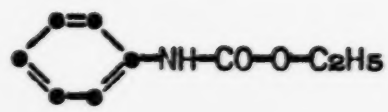
EXAMPLE 5 - ANSWER 2

REGISTRY NO. = 101,995

$C_9H_{11}NO_2$

Preferred Name: Carbanilic acid, ethyl ester

SOCMA-590l
C62:1995C
CZ-2653
C63:4490g
C63:4490g
C63:7588d
C63:11291a



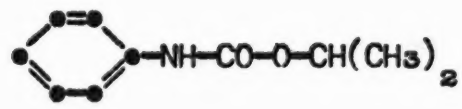
EXAMPLE 5 - ANSWER 3

REGISTRY NO. = 122,429

$C_{10}H_{13}NO_2$

Preferred Name: Carbanilic acid, isopropyl ester

socma-252i
c62:21908
c62:3332h
c62:3333c
c62:3334c
c62:5819h
c62:7041f
c62:8329f
c63:6255d
c63:7588d
c63:8973b
c63:11291a



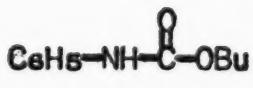
EXAMPLE 5 - ANSWER 4

REGISTRY NO. = 1,538,745

$C_{11}H_{15}NO_2$

Preferred Name: Carbanilic acid, butyl ester

c62:p161218
c63:07588d
c63:00451f
c63:07588d



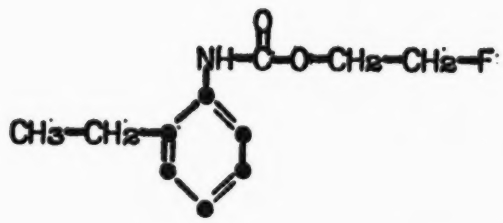
EXAMPLE 5 - ANSWER 5

REGISTRY NO. = 1,542,489

$C_{11}H_{14}FNO_2$

Preferred Name: Carbanilic acid, o-ethyl-, 2-fluoroethyl ester

C52:3691h



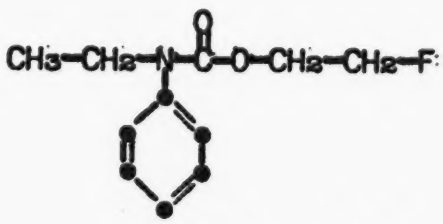
EXAMPLE 5 - ANSWER 6

REGISTRY NO. = 1,542,490

$C_{11}H_{14}FNO_2$

Preferred Name: Carbanilic acid, N-ethyl-, 2-fluoroethyl ester

C53:1198b



F-NSF
SE

FROM:

ERIC FACILITY
SUITE 601
1735 EYE STREET, N. W.
WASHINGTON, D. C. 20006